Benefit-Risk Assessment in Drug Development

Sarac, Sinan

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Benefit-Risk Assessment in Drug Development

PhD thesis by Sinan B. Sarac

Supervisors:

Professor Emeritus Erik Mosekilde, DTU Physics

Affiliate Professor Morten Colding-Jørgensen, Quantitative Clinical Pharmacology, Novo Nordisk A/S

Affiliate Professor Steffen Thirstrup, Danish Health and Medicines Authority

Examiners:

Professor Per Helboe, University of Copenhagen, Denmark

Professor Stuart Walker, University of Wales, Cardiff, UK

Professor Ole Lund, DTU Systems Biology

Chairman:

Associate Professor Jørn Bindslev Hansen, DTU Physics

Submitted:

December 23rd 2011

Defended:

April 16th 2012

Acknowledgments

This thesis presents the work that I have carried out during my PhD studies at the Technical University of Denmark, the Danish Medicines Agency and Novo Nordisk A/S, 2008-2011. The work has been financially supported by the Danish Ministry of Science, Technology and Innovation and Corporate Research Affairs, Novo Nordisk A/S. The studies have been supervised by Professor Morten Colding-Jørgensen, Novo Nordisk A/S, Professor Erik Mosekilde, Department of Physics, Technical University of Denmark, and Professor Steffen Thirstrup, Head of Licensing Department, Danish Medicines Agency. The views and opinions expressed in this paper are those of the author and should not be attributed to Novo Nordisk A/S, Danish Medicines Agency or Technical University of Denmark.

I would like to thank my dear wife and life companion, Nuray B. Sarac, for her encouragement, support and love. This thesis is only a reality because of her. My dear children, Elias, Alan and Amalia, always kept me busy in my spare time. They are thanked for putting life into perspective and letting me remember that there is more to life than research and work.

A number of individuals have had a scientific impact on my work. First of all, I would like to thank my friend, colleague and “younger brother”, Christian Hove Rasmussen. It has been a pleasure working with him, and he is thanked for inspiring discussions and invaluable input to this thesis. Morten Arendt Rasmussen and Christine E. Hallgreen are also thanked for enlightening conversations.

Finally, I would like to thank my supervisors. Morten Colding-Jørgensen is acknowledged for his continuous guidance and supervision as well as his never-ending encouragement and support. Erik Mosekilde is thanked for his supervision, support and encouragement. Steffen Thirstrup is thanked for his advice and support. Steffen always found new solutions and opportunities when most needed.

Sinan B. Sarac
Copenhagen, December 2011

Table of content

List of Abbreviations

CA	Conjoint Analysis
CER	Comparative Effectiveness Research
CHMP	Committee for Medicinal Products for Human Use
CIOMS	Council for International Organizations of Medical Sciences
CIRS	Centre for Innovation in Regulatory Science
DALY	Disability-Adjusted Life Years
D80 AR	Day 80 Assessment Report
EMA	European Medicines Agency
FDA	US Food and Drug Administration
IMI	Innovative Medicines Initiative
MCDA	Multi-Criteria Decision Analysis
MCE	Minimum clinical efficacy
MDR	Multifactor dimensionality reduction method
NNH	Number Needed to Harm
NNT	Number Needed to Treat
PhRMA	Pharmaceutical Research and Manufacturers of America
PROTECT	Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium
QALY	Quality-Adjusted Life Years
Q-TWIST	Quality-Adjusted Time Without Symptoms or Toxicity
RV-NNT/H	Relative Value-adjusted Number Needed to Treat/Harm
TTD	Time to death
TTR	Time to relapse
5-FU	5-Fluorouracil

English summary

This thesis covers the development, testing and use of an eight-step structured method for data-driven benefit-risk assessment. The aim of this thesis was to create a tailored method for the assessment of clinical data. The focus has been on three major aspects: (i) transparency, (ii) clinical significance, and (iii) visualisation.

A simple preliminary method was created and tested in a pilot study and was presented at an internal workshop in Novo Nordisk A/S, where the given project team was invited. Input from the project group was used to adjust and optimise the method, which was then tested in a new pilot study, and the results were presented at a new workshop. In total, four pilot studies and internal workshops were conducted. The method was therefore developed in an iterative fashion.

The method involves eight successive steps: 1) establishment of the decision context, 2) identification of benefit and risk criteria, 3) weighting, 4) scoring, 5) evaluation of uncertainty, 6) calculation of weighted scores, 7) visualisation, and 8) discussion and formulation of an overall conclusion. In order to reduce the impact of subjective judgments, scores are assigned to each criterion on the basis of objective information (data) wherever possible.

The method is comprehensive and supported by a qualitative framework with built-in quantitative measures. However, at the same time the method is transparent in the sense that all assumptions made in the various steps of the assessment are clearly expressed all the way to the final decision. This is important both to avoid that unreported biases and feedback distort the assessment, and to make it possible for the industrial partner and the regulatory agency to compare the results of their evaluation on a point-by-point basis. The qualitative framework ensures a structured approach to the assessment and a transparent communication of the results.

Clinical significance and relevance of data is defined in qualitative ways, but captured by quantitative measures, enabling a discussion of the clinical relevance of data. This has

opened a new dimension in the discussions related to data. Different approaches to different types of data have been developed, tested and used.

Standardised diagrams for the visualisation of results from the assessment have been established, and different diagrams have been developed for different scenarios. For the visualisation of results from single and/or multiple similar trial assessments, *tornado-like* diagrams were designed. For dose-finding studies and multiple non-similar trial assessments, *matrix* diagrams have been developed and for events *scoring* tables have been set up. All of these different visualisation techniques enable a transparent communication of assumptions and decisions made in the assessment.

The method was successfully applied to three different cases: (i) colorectal cancer, (ii) schizophrenia, and (iii) telithromycin (Ketek®), and demonstrated its general applicability.

Dansk resumé / Danish summary

Denne afhandling omhandler udviklingen, afprøvningen og brugen af en metode for data-dreven benefit-risk evaluering. Metoden blev udviklet som iterativ proces og er skræddersyet til benefit-risk evaluering af kliniske data. Fokus har været på tre afgørende aspekter: (i) gennemsigtighed, (ii) klinisk signifikans og (iii) visualisering.

Metoden er vidtfavnende og er støttet af et overordnet kvalitativt skelet med indbyggede kvantitative metoder, men samtidig er strukturen transparent, idet alle antagelser foretaget i forbindelse med en given evaluering, tydeligt kommer til udtryk i den endelige beslutning. Dette er særlig vigtigt både for at undgå at ikke-rapporteret bias og feedback fordrejer en evaluering, hvilket kan gøre det umuligt for den industrielle partner og den involverede regulatoriske myndighed at sammenligne resultaterne af deres evalueringer skridt for skridt. Det kvalitative skelet sikrer en struktureret tilgang til alle evalueringer og en transparent kommunikation af de opnåede resultater.

Klinisk signifikans og relevans af data er defineret via kvalitative kriterier, men bliver tilvejebragt ved hjælp af kvantitative metoder, hvilket muliggør en diskussion af klinisk relevans af data. Dette har åbnet en ny dimension i diskussioner relateret til data. Forskellige metoder til forskellige typer af data er blevet udviklet, testet og brugt.

Der er blevet formuleret standardiserede diagrammer for visualiseringen af resultaterne fra evalueringer og et sæt af diagrammer er blevet udviklet til forskellige scenarier. Til visualisering af resultater fra enkelt og/eller multiple lignende studie evalueringer er der designet tornado-lignende diagrammer. For dosis studier og multiple disparente studie evalueringer er der blevet udviklet matrix diagrammer. For begivenheder (bivirkninger, responder rate, m.m.) er der blevet udviklet skorings tabeller. Disse visualiserings metoder muliggør en transparent kommunikation af antagelser og beslutninger i evalueringen.

List of publications

1. Sarac, S.B., Rasmussen, C.H., Rasmussen, M.A., Hallgreen, C.E., Søeborg, T., Colding-Jørgensen, M., Christensen, P.K., Thirstrup, S., Mosekilde, E. *A Comprehensive Approach to Benefit-Risk Assessment in Drug Development.* [In Press]. Article first published online: 17 MAR 2012 | DOI: 10.1111/j.1742-7843.2012.00871.x. Basic & Clinical Pharmacology & Toxicology (BCPT).

2. Sarac, S.B., Rasmussen, C.H., Afzal, S., Thirstrup, S., Colding-Jørgensen, M., Poulsen, H.E., Mosekilde, E. *Data-driven assessment of the association of polymorphisms in 5-Fluorouracil metabolism genes with outcome in adjuvant treatment of colorectal cancer.* [In Press]. Article first published online: 16 APR 2012 | DOI: 10.1111/j.1742-7843.2012.00885.x. Basic & Clinical Pharmacology & Toxicology (BCPT).

3. Shahrul Mt-Isa, Nan Wang, Christine E. Hallgreen, Torbjörn Callréus, Georgy Genov, Ian Hirsch, Steve Hobbiger, Kimberley S. Hockley, Davide Luciani, Lawrence D. Phillips, George Quartey, **Sinan B. Sarac**, Isabelle Stoeckert, Alain Micaleff, Deborah Ashby, Ioanna Tzoulaki. *Benefit-risk Integration and Representation: A Systematic Review and Classification of Methodologies for Benefit-Risk Decision-Making in Medicines.* [Prepared on behalf of IMI PROTECT].

4. Gesche Jürgens et al. Health Technology Assessment of genotyping in patients with schizophrenia: *"Do individual dosing supported by genotyping improve the medicinal treatment with antipsychotics?* [Prepared for the National Board of Health, Denmark]. (I have contributed to the part concerning the clinical benefit-risk assessment of data). [DRAFTED].

1 Introduction

1.1 Motivation

Benefit-risk analysis requires a significant amount of time and resources from regulatory agencies and pharmaceutical companies. However, the purpose of this activity often appears to be inadequately defined and the outcome often seems to be a result of personal experience and expertise[1]. A number of recent cases of critical safety issues, e.g. Vioxx in connection with cardiovascular deaths[2] and Avandia and the higher risk of myocardial infarction[3], have caused increased risk awareness and risk aversion from regulatory agencies. As a result, the development cost of a new drug now often exceeds more than 1 billion USD and the development process takes about 12-14 years[4]. Nevertheless, the benefit-risk considerations regarding whether or not to submit or grant a licence to a drug are still not performed according to a well-established framework[5].

In 1998 the Council for International Organizations of Medical Sciences (CIOMS)[1] expressed that the terms benefit-risk assessment, benefit-risk ratio, etc., in spite of common use, were inadequately defined. Although regulators and the industry often made benefit-risk analysis, it was concluded that there were no generally acceptable methods or models. However, such assessments and analysis were often demanded or requested by regulatory agencies without any formal guideline or tools being provided.

Several years later, both the US Food and Drug Administration (FDA)[6-8] and the European Medicines Agency (EMA)[9] emphasised the need for more structured and transparent benefit-risk assessments in connection with the evaluation of new pharmaceutical products. Both agencies have on-going activities aiming at developing a standardised framework for benefit-risk assessment[10]. As a response to the CIOMS IV report[1], several pharmaceutical companies, institutions, etc. began developing methods and frameworks for benefit-risk assessment. Prior to 1998 and the CIOMS IV report, there were many attempts to assess specific aspects of the benefit-risk profile of drugs[10;11]. After the CIOMS IV report, there has been continuously increasing activity within the field: some

general benefit-risk assessment methods have been developed and several reviews have been published on this topic[1;9-16].

Many approaches have been based on Multi-Criteria Decision Analysis (MCDA)[5;17-20], which is well-known and widely used in other areas of society, e.g. political decision making[21;22]. In a report from 2008 on benefit-risk assessment models and methods by the Committee for Medicinal Products for Human Use (CHMP)[9], MCDA is also mentioned as a possible and promising way to cope with multiple objectives and criteria. Nevertheless, CHMP expresses concerns such as the use of linear combinations of single values and the lack of attention to uncertainty. The method focuses on decision making, where overall values are compared, rather than looking at the elements individually. Based on a review of some existing methods, CHMP concludes that there is a need for a tailored method for medical benefit-risk assessment in relation to drugs. In 2009 the Benefit-Risk Methodology Project began at the EMA[9;10]. The aim is to develop a structured and transparent approach to benefit-risk assessment. The project is still in progress and no final conclusions have been made.

FDA is currently working towards a qualitative approach that should bring transparency and structure to the decision on a benefit-risk profile[23;24]. For almost a decade, the Centre for Innovation in Regulatory Science (CIRS) has been dedicated to the development of a simple and structured approach to benefit-risk assessment. Since 2008, a consortium consisting of Health Canada, Therapeutics and Goods Administration (TGA, Australia), Health Services Authority (HAS, Singapore) and Swissmedic has worked in close collaboration with CIRS on the development of a common approach to benefit-risk assessment.

The aim of all these activities is to enable the decision makers to comprehend and keep in mind all relevant aspects of the assessment to provide the basis for a more justified decision. Every medicine has both risks and benefits and the aim is to put these benefits and risks into perspective and communicate the decision in transparent ways. Benefits and risks are recorded differently in trials using different methods and statistical tools and consequently

there is different evidence to support the conclusions. Many benefit-risk assessments are subjective in their handling of data from clinical trials[5;9]. They depend on input from key individuals at crucial stages of the decision-making process. The assessment can therefore differ substantially between companies and regulatory agencies, as described by Boudes[25]. Boudes concludes that “While statistically significant results are necessary, they are not sufficient, and the clinical relevance of the data should be explained”. This appears to be the main reason for many drugs not getting approved by regulatory agencies. There is simply a lack of discussion of the clinical relevance of data.

Research within psychology and the cognitive functions of the human mind shows that the amount of information comprehended at any given point of time rarely exceeds 7±2 “chunks” of information[26]. When faced with multiple “chunks” or an increasing amount of information, one unconsciously seeks a way to comprehend the information by sorting it into various clusters. However, there is a natural limit to the size of these clusters. It is obvious that the amount of information in a clinical trial is overwhelming and it is further increased by multiple clinical trials. How can a person assessing these data comprehend all these pieces of information, which are different in both quality and nature, at one and the same time?

When faced with multiple clusters of information, e.g. benefits and risks, we intuitively decide to focus on the most important benefits and risks. The decision is then based on a balance between few aspects of, e.g. a new medicine. Is this correct and rational? The increasing attention to the field of benefit-risk assessment within regulatory agencies as well as the pharmaceutical industry answers this question: No, it is neither correct nor rational to focus on a few criteria when deciding the fate of a new drug. A more comprehensive approach is much needed.

At present, though, to our knowledge, the regulatory agencies have not yet provided the pharmaceutical industry with clear and detailed guidelines for the preparation of assessments, and a number of different benefit-risk assessment methods are currently in use. However, none of these are generally accepted and used.

1.2 Thesis objectives and outline

Transparency and simplicity are key elements in a benefit-risk assessment, enabling any stakeholders to grasp and understand the reasoning behind a decision. The overall objective of this thesis is devoted to the development of methods that can enhance these facets.

Pharmaceutical companies invest millions of USD into clinical drug development projects and thousands of patients are enrolled in their clinical programmes even though the failure rate of a drug candidate is extremely high. A new methodology that can predict or minimise the risk and reduce failure rates during clinical drug development will therefore have an enormous impact both commercially and with respect to patient safety. The present project aims to develop a general benefit-risk assessment method that can be used in drug development, during marketing authorisation and in postmarketing surveillance.

There is considerable uncertainty about the best model and method to determine the benefit-risk balance and to express the research results in such a manner that they assist decision making. The main reason for this uncertainty is that many prior attempts have been based on theories about decision making and value trees combined with subjective voting among groups of decision makers to score the different criteria and weights without using data.

The main part of this project is to develop a qualitative framework to assess variables encountered in clinical trials, e.g. disease progression, treatment outcome, adverse effects etc., and to compare different qualities of variables, e.g. patient compliance against treatment effect, different benefits or adverse effects for different subgroups, etc. The aim of this thesis is to develop a simple, structured and transparent approach to the benefit-risk assessment of medicines and treatments with a focus on the clinical significance and relevance of data and how these results can then be communicated in a simple and transparent fashion, once the subjective input to the assessment is justified and quantified. It is important to emphasise that the aim is not to develop a decision model, but rather an approach that supports the intellectual decision-making process and increases the transparency of the assessment at the same time. This is an important feature when communicating results of the assessment.

The method will be built in an iterative fashion and will focus on the clinical stages of drug development, mainly phase II and III.

The first part of this thesis is concerned with the evaluation of existing methods and tools for the benefit-risk assessment of data. Methods will be divided into main groups and the most influential methods will be described and evaluated. The pros and cons will be discussed. The aim is to investigate what is missing and to identify aspects that should be included in a novel benefit-risk assessment methodology. Current initiatives will be evaluated and their similarities and differences will be discussed. Major aspects and ideas that should be part of a benefit-risk assessment will be identified, evaluated and put into perspective.

In the second part of this thesis, the focus will be on the clinical significance and relevance of data and how this can be defined and quantified. Often, statistical significance is incorrectly used to claim clinical significance. Differences between statistical significance and clinical significance will be discussed. Past definitions of clinical significance in the literature will be discussed and evaluated. Furthermore, it will be discussed how clinical significance can be incorporated in a benefit-risk assessment and how it can be communicated in a simple and transparent way. The development of a novel methodology and visualisation tool will be described as well as how this relates to existing methods and tools. The real-life testing of these ideas in four workshops and the learning will be elaborated, providing the reader with a clear understanding of the iterative development of the proposed methodology.

The final and third part is concerned with the results gained from the use of the method. A step-by-step presentation of publishable data is given. This final part is also concerned with discussions with the Danish Medicines Agency and their views on benefit-risk assessment in general and on the proposed methodology in particular.

1.3 Main results

The method has been developed in an iterative process and tested on four different drugs in Novo Nordisk A/S (confidential). The work resulted in a scientific paper describing the method step by step, using hypothetical examples to demonstrate all the qualities. The method constitutes eight steps, focusing on creating transparency in the assessment by always justifying choices, minimising correlations in data, and avoiding summation of data, e.g. benefit:risk ratio, etc. There is focus on clinical significance and relevance, which is defined as the proportion of patients experiencing a defined treatment effect. Different visualisation techniques are used to communicate results.

The method has been used on several occasions:

- Benefit-risk assessment of data from a schizophrenia database (Bispebjerg Hospital, Denmark).
- Benefit-risk assessment of data from a colorectal database (Copenhagen University Hospital, Denmark). The method clearly demonstrated that there were clinically significant results for patients with a specific combination of polymorphisms with regard to time-to-death, fatigue and mucositis/stomatitis. This study indicated that genotyping in colorectal cancer could aid in individualised monitoring and treatment of patients. The work resulted in a scientific paper.
- Benefit-risk assessment of Ketek® is based on data from the EPAR from the European Medicines Agency. The assessment was conducted in cooperation with the Innovative Medicines Initiative (IMI) project *Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consort* (PROTECT). This assessment was conducted not to challenge the conclusions in the EPAR made by the EMA, but to investigate the implications of the application of a structured approach. The method demonstrated the main differences between Ketek and existing comparators. The results were clearly communicated and visualised.

The main advantages of this method are:

1. The focus on transparency by using a successive stepwise approach that diminishes unintended bias and feedback, by justifying every choice in the assessment, by ensuring thorough discussion of choices, etc.
2. The focus on clinical significance and relevance of data, which add a new dimension to the discussions related to data and the statistical analysis made on data.
3. The focus on visualisation of the results. There is an enormous difference in the comprehension of results by using shapes and colours in standardised diagrams. Several different standardised diagrams were created for different scenarios.
4. The flat structure, where no overall, e.g. benefit and risk score, benefit-risk ratio, etc. are calculated. The effects of correlations are diminished and controlled.

I have, as part of a group, on behalf of PROTECT, prepared a comprehensive review and classification of methodologies for benefit-risk decision making in medicines. The review is in its final stages and is expected to be published in the near future.

The method has been presented at more than 15 internal meetings/conferences in Novo Nordisk A/S. On a national scale, the method has been presented by invitation at several meetings and conferences, e.g. the Danish Medicines Agency, the Danish Society for Pharmacology, Bispebjerg Hospital (Department of Clinical Pharmacology), etc.

For an international forum, the method has been presented by invitation at the Centre for Innovation in Regulatory Science (CIRS), formerly known as *CMR International Institute for Regulatory Science*, Washington, USA, and at DIA's 47th annual meeting in Chicago, USA.

2 Basic methods

2.1 Defining benefits and risks: a common language

Only the approaches of the two major regulatory agencies (EMA and FDA) will be described here. Regarding EMA, several main documents are identified, e.g. the *"Day 80 Assessment Report (D80 AR): overview and list of questions"*[27] and the *"D80 AR: clinical aspects"*[28]. In the "D80 AR: overview and list of questions", there is a section specifically dedicated to benefit-risk assessment. The purpose of this section is to provide a snapshot of the key benefits and risks, of the strengths of evidence and the limitations of data. There are no specified guidelines for the assessor to conduct any formal benefit-risk assessment; it is therefore up to the assessor to use any method, model, tool and/or framework that he/she finds suitable. The benefit-risk assessments are consequently influenced by personal experience and expertise of the individual assessors.

Decisions are made in groups and not always in consensus, and there is a CHMP guideline for the voting system[29]. Many terms are not defined and questions remain unanswered, and the assessments may differ. Therefore, consistency in the decision process may vary from assessment to assessment. As discussed later in this chapter, these shortcomings have been acknowledged by the EMA and consequently the Benefit-Risk Methodology Project has been initiated.

From FDA, many documents describing the current recommendations and guidance can also be identified, i.e. *"FDA Guidance for Industry: Premarketing Risk Assessment (Premarketing Guidance)"*[30], *"FDA Guidance for Industry: Development and Use of Risk Minimization Action Plans"*[31], *"FDA Guidance for Industry: Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment"*[32], *"FDA Good Reviewer Practice: Clinical Safety Review of an NDA or BLA"*[33]. These documents cover pre- and post-marketing phases of drug development and approval. Similar to the documents from EMA, phrases and terms like "benefit-risk balance", etc. are used without any definition. There is no

explicit guidance on how benefit-risk assessment should be conducted. Thus, the assessment is qualitative and is again influenced by personal expertise and experience.

Both FDA and EMA have realised that there is a need to improve the quality of decisions. For several years, a group of specialists within the FDA has worked on a qualitative framework for benefit-risk assessment[10]. In 2009 the EMA initiated the "Benefit-Risk Methodology Project" and have been very open about their initiative. The first work package set out to investigate how different member countries defined benefit and risk and how benefit-risk assessments were performed at different agencies[34]. In total, 55 people involved in benefit-risk decision making from the five participating agencies (France, The Netherlands, Spain, Sweden, and the UK) and the Paul-Erlich-Institute (Germany) were interviewed.

To enable a benefit-risk assessment in consensus, there must be a common understanding of the terms benefit and risk. The results showed that both terms were widely used, but without any kind of common definition or understanding. Deciding the benefit-risk balance was generally considered as the most difficult part of the assessment. The most difficult assessment involved the cases with a high degree of uncertainty involved, i.e. new classes of drugs and drugs belonging to the therapeutic area of oncology.

This may explain why there are differences of opinion on the same dataset, maybe not within the same agency, but definitely between agencies. The conclusion from the report was that there was no clear understanding or definition of these terms and the benefit-risk methodology project team decided to avoid using them. Instead, the terms "favourable effects" and "unfavourable effects" were adopted and defined as follows:

Favourable effects are any beneficial effects for the target population (often referred to as "benefits" or "clinical benefits") that are associated with the product.

Unfavourable effects are any detrimental effects (often referred to as risks, harms, and hazards both known and unknown) that can be attributed to the product or that are otherwise of concern for their undesirable effect on patients' health, public health, or the environment.

In this thesis, benefits and risks will be defined in the following paragraphs. EMA's definition of favourable and unfavourable effects will be used and further elaborated upon:

Benefit criteria are defined as any beneficial effects for the target population that are associated with the product, and all potential or reported benefits. Potential benefits are based on non-clinical findings, mechanisms of action, beneficial effects of the pharmacological class, off-label use, and presumed long-term benefits not yet fully explored. The benefit criteria are based on objective measurements (e.g. biomarkers) and semi-objective measurements (e.g. convenience, quality of life).

Risk criteria are defined as any detrimental effects (often referred to as harms, and hazards both known and unknown) that can be attributed to the product or that are otherwise of concern for their undesirable effect on patients' health, public health, or the environment. Furthermore, all potential or reported adverse events, including both serious and non-serious events, are defined as risks. Potential risks are based on non-clinical safety findings, mechanisms of action, adverse events of the pharmacological class, drug-drug interaction, risk of overdosing and off-label use and long-term safety or rare adverse events not yet fully explored.

The report from the first work package of EMA's benefit-risk methodology project indicated that transparency and consistency in decisions must be the main pillars whereupon future decisions are made. Transparency is enhanced by common language, while consistency is enhanced by a standardised framework, method, model or tool. Transparency and consistency are both in the interest of the regulatory agencies and the pharmaceutical industry. The main responsibilities of regulatory agencies are to protect the public from harm. Drugs that are safe are approved, but the decision is based on data, where there is often great uncertainty regarding the safety profile of the drug. The efficacy of the drug is usually well established at the time of approval.

The pharmaceutical industry, on the other hand, is interested in a consistent methodology or guideline, because of the enormous economical investments made to bring a drug to the market. It is therefore critical that data is considered in a standardised way, avoiding “gut-feelings” that might cause insecurity and result in costing the firm millions or billions of dollars, and maybe depriving the public of a much needed treatment.

Regulatory agencies often operate from a precautionary principle, which first appeared in its modern form in the 1970s in Germany. There are numerous examples of the use of this principle by, e.g. regulatory agencies and legislative organs. The precautionary principle is intended as a supportive decision-making tool in decisions that are surrounded by risk and uncertainty[35]. However, the principle does not provide a method/tool or framework on how to guide a decision under uncertainty.

The precautionary principle states that when a decision is made with a suspected risk of causing harm, the decision maker is responsible for any consequence of the decision. This has therefore forced regulatory agencies to take a very defensive position to protect the public. Once more and more scientific information provides evidence that no harm will come to the public, these restrictive actions are relaxed.

The quantification of risk is difficult because of the nature of these data. Data is often sparse and sporadic. The classical statistical inference approach will often have shortcomings in the handling of this type of data. There have been attempts to implement a probabilistic (Bayesian) approach to the precautionary principle. The end result was more or less Bayesian decision making, with the definition of prior probabilities and calculation of posterior probabilities[36]. However, this method is restricted to risk management.

The precautionary principle is restrictive and could therefore hamper drug discovery and development, because of an overwhelming focus on risk. The principle is too vague and provides no framework or method/tool that can aid decision making. There is a lack of focus on benefits. However, the precautionary principle could be useful in aiding discussions to clarify some aspects of the decision context.

The aim of a standardised methodology must be to put different benefits and risks into perspective and accommodate for uncertainty in the data, visualising the results of the assessment. This will enable regulatory agencies to protect the public and make consistent decisions, allowing the industry to be confident that their drug is protected from the unreasonable insecurity and “gut feelings” that some regulatory assessors might have.

Therefore, a methodology for drug development should allow different types of data to be brought onto the same level. This will enable a direct comparison of benefits and risks, hereby putting them into perspective. To enhance the consistency of decisions throughout the drug development process and the communication of decisions made at critical points, results of the assessment must somehow be compromised for display, but without losing valuable information.

2.2 Evaluation of existing methods and current initiatives

There are many methods and tools that can be used for benefit-risk assessment of interventions. Some are more comprehensive than others and some are quantitative, while others are qualitative. The different methods/models/tools can be clustered into five different categories:

1. Metrics
2. Quality of life measurements
3. Survey methods
4. Early benefit-risk methods
5. Comprehensive benefit-risk methods

The different methods/models/tools will not all be described; instead a general discussion of the category will be provided, while only the main methodologies within each category will be elaborated. Classical statistical inference tests will not be evaluated in a comprehensive way, since this would be out of the scope of this thesis. Each category will be discussed,

evaluated and its advantages and disadvantages highlighted. The purpose and aim of this section is to clarify which aspects enhance the transparency, the objectivity and consistency of decisions and the communication of the results and decisions of the assessment.

The focus on transparency and a structured approach to benefit and risk evaluation was initiated in the 1970s by Tallarida and co-workers[37], who, as pioneers, defined numerical scales for the severity of adverse events and weighted them against the primary illness being treated. They concluded that a numerical scale was superior to non-numerical scales of severity, like “mild”, “moderate” and “severe”. This attempt to quantify the qualitative judgement behind medical decisions was, as we will see later in this section, far ahead of its time.

2.2.1 Metrics

In this section, the focus will be on methods like Number Needed to Treat (NNT)[38], Number Needed to Harm (NNH)[38], Relative Value-adjusted NNH (RV-NNH)[39;40], Minimum clinical efficacy (MCE)[39;40], Incremental Net Benefit (INB)[5;41;42], Adverse event-adjusted NNT (AE-NNT)[44] and Utility- and time-adjusted NNT (UT-NNT)[43]. These methods represent an attempt to transform clinical results into a numerical output that can be evaluated in a qualitative manner.

All of the abovementioned methods are related to or originate from the original work of A. Laupacis, D.L. Sackett and R.S. Roberts (1988), who were frustrated over the lack of tools to measure and compare the benefits and risks of, e.g. different treatments. They argued that the reciprocal of absolute risk reduction (ARR) may be a highly useful measure for clinicians in communicating the benefits and risk of different treatments to the patient:

$$ARR = p_a - p_b$$

and

$$NNT = \frac{1}{p_a - p_b},$$

where $p_a > p_b$ and are the probabilities for a given event related to treatments *a* and *b*. The events can be responder rates that represent success for two given treatments.

NNT is the number of patients that must be treated in order to prevent one adverse event or to observe a positive event, and NNT usually refers to a therapeutic intervention. NNH (Number Needed to Harm) is similar to NNT, but contrary to NNT, NNH refers to risk. For the rest of this chapter, we will concentrate on NNT, because whatever is said about NNT will apply to NNH as well. NNT is one of the first attempts of translational medicine aiming at a benefit-risk evaluation and several advantages are proposed[38;44]:

1. NNT is simple to calculate
2. As an integer it is simple to remember
3. It is more readily understood than its alternatives
4. It can be used to describe the harm as well as the benefit of therapy and other clinical issues

However, several studies have shown that the above-listed arguments for the use of NNT are not to be taken literally. Patients and physicians are willing to accept a treatment regardless of NNT as long as there are no major risks associated. Moreover, both patients and physicians tend to consider therapy like a lottery, where the probability of receiving the benefit of treatment is 1/NNT. This means that NNT is not easily understood and is not intuitive. Another drawback is the question of whether one single measure for benefit and risk can represent all the necessary information and justify an overall benefit-risk evaluation[45].

Several other aspects of NNT have been criticised. The estimate is biased, because there is no consideration of variability. It is difficult to calculate CI (confidence intervals) and they are not very helpful[39;44]. CI of NNT can be calculated as the inverse of the lower and upper

limits of the CI for ARR, e.g. if the risk difference in a trial is 10% with a 95% CI from 5% to 15%, the NNT is 1/0.1 = 10 and the 95% CI for the NNT is 6.7 to 20 (1/0.15 to 1/0.05). If ARR is not significant and the CI includes 0, problems arise in the calculation of CI for NNT, e.g. if ARR is 10% with a wider 95% CI, -5% to 25%, then the 95% CI of NNT is -20 to 4. There are two problems associated with this estimate: 1) NNT can only be positive, and 2) the CI does not include the estimate of NNT of 10. When the CI of ARR includes 0, the CI of NNT becomes non-intuitive, because NNT for 0 is infinity (∞)[46].

Like other statistical methods, NNT/NNH are often subject to misuse and misinterpretation[47;48]. This, by itself, is obviously not a reason to discard NNT/NNH. It is important to bear in mind that NNT/NNH is a descriptive method for presenting results, not of data analysis[49].

Even though NNT/NNH are simple and practical in their use, they cannot be combined to determine the benefit-risk balance of drugs, because there is no account of clinical significance[10]. Holden suggested that relative utility values and minimum clinical efficacy analysis should be incorporated in NNT/NNH calculations[10]. The lack of quantitative methods for benefit-risk assessment of drugs led Holden and co-workers to propose different methods for more quantitative assessments. They point out that the vast majority of benefit-risk assessments are subjective and simplistic descriptions or a review of, e.g. clinical trial results, with no quantification at all[11;39].

They recommend and propose that relative value-adjusted NNT (RV-NNT) and relative value-adjusted minimum clinical efficacy (MCE) can be used to enable a more quantitative benefit-risk assessment. RV is calculated as follows:

$$RV = \frac{1 - utility.of.AE}{1 - utility.of.disease},$$

where utility is defined as a numerical value based on patients' preference for specific outcomes of a given treatment and AE is adverse events. RV can be interpreted as the value of avoiding an AE relative to avoiding the disease of interest or target event.

$$RV - NNT = \frac{1}{(p_a - p_b) \times RV},$$

where $p_a > p_b$ and are the probabilities for a given event related to treatments *a* and *b*. A treatment is warranted when NNT < NNH, which takes into account all AEs of interest and their relative utility values compared with the utility value of the disease of interest. An adjusted NNH for multiple adverse events can also be calculated. All the pros and cons associated with NNT/NNH also apply to RV-NNT/-NNH.

MCE resembles RV-NNT/-NNH, but tries to incorporate a calculated weighting of potential benefits and risks of a given treatment,

$$E(T_1) \geq E(T_2) + \frac{R(T_1) - R(T_2)}{p(T_0)},$$

where,

$$E(T_1) = \frac{p(T_0) - p(T_1)}{p(T_0)}$$

and

$$E(T_2) = \frac{p(T_0) - p(T_2)}{p(T_0)},$$

where T_1 and T_2 are defined as treatment 1 and 2, $E(T_1)$ and $E(T_2)$ are defined as the efficacy of T_1 and T_2 , $R(T_1)$ and $R(T_2)$ are defined as the risk of a given AE when treated with T_1 and T_2, and $p(T_0)$ is defined as the risk of disease of interest in the untreated/placebo group. From the above equation, we can conclude that for new treatment to be better than an

existing one, the efficacy of the new treatment must be at least the same as the efficacy of the present treatment. This method can also be extended to include multiple AEs.

RV-NNH and MCE have an advantage over NNH/NNT in the incorporation of some kind of utility/preference and weighting. This is a key element in the direct comparison of benefits and risks. However, neither method defines clinical significance. In MCE, a new treatment only has to be the same or slightly better to be worth considering. The next obvious question is: how much better should a new treatment be, to be perceived as clinically significant and relevant? This question remains unanswered.

There is no real value in a benefit-risk setting in calculating one single number. Information is lost every time values are added, multiplied, etc. In MCE, prior probabilities are applied and there is always a risk that these can be incorrect, which leads to useless results.

2.2.2 Quality of life measurements

This section will cover measures like quality-adjusted life years (QALY)[10;50-53], disability-adjusted life years (DALY)[54], health-adjusted life expectancy (HALE)[55] and quality-adjusted time without symptoms or toxicity (Q-TWIST)[56].

These measures stem from a frustration over the lack of attention to, e.g. disability, quality of life, health, etc. In 1990 the World Health Organization (WHO) wanted to express the burden of disease with measures other than mortality and morbidity. This request was answered by Prof. Murray at Harvard University, who presented the WHO with the DALY. DALY was adopted by WHO in 2000. DALY combines the potential life years lost due to premature death with equivalent years of “healthy” life lost due to poor health or disability. It combines mortality with morbidity[54].

Q-TWIST was originally developed for the evaluation of different treatment regimes in oncology, but has since been applied to other fields, e.g. HIV treatment[57]. The method expresses the need for qualitative measures for the quality of life for cancer patients treated

with chemotherapy. It reflects the dilemma between extending a cancer patient's life, but also decreasing the quality of life at the same time.

The method can be divided into three main steps: 1) the quality-of-life-oriented endpoints are defined, 2) the survival time is partitioned, and 3) the different treatments are compared regarding the quality-adjusted survival. Survival time is divided into "time with relapse" (REL), "time with toxicity" (TOX) and time without symptoms and toxicity (TWIST). Based on these estimates, a quality-adjusted survival model can be constructed using utilities for periods with toxic effects (u_{tox}) and periods of relapse (u_{rel}). A scale from 0 to 1 can be used, e.g. $u_{tox} = 0.8$ and $u_{rel} = 0.5$, assuming that patients and physicians are more willing to tolerate toxic effects than disease relapse.

In reality, this means that weights are assigned. TWIST is assigned the weight of 1 and death 0. The Q-TWIST score can then be calculated as follows:

$$Q - TWIST = (u_{tox} \times TOX) + TWIST + (u_{rel} \times REL)$$

Different treatment regimens can be compared based on their Q-TWIST score[56]. However, the interpretation of these results is not straightforward. Although Q-TWIST is frequently used in the evaluation of cancer clinical trials, there is no clear definition of what defines a clinically relevant difference. Revicki and co-workers performed an extensive review based on cancer clinical trials that used Q-TWIST in the analysis of data; they recommend that a clinical relevant difference for Q-TWIST is 10% of overall survival and differences of 15% are considered clearly clinically important[58]. Figure 1 shows how Q-TWIST is used in a clinical setting.

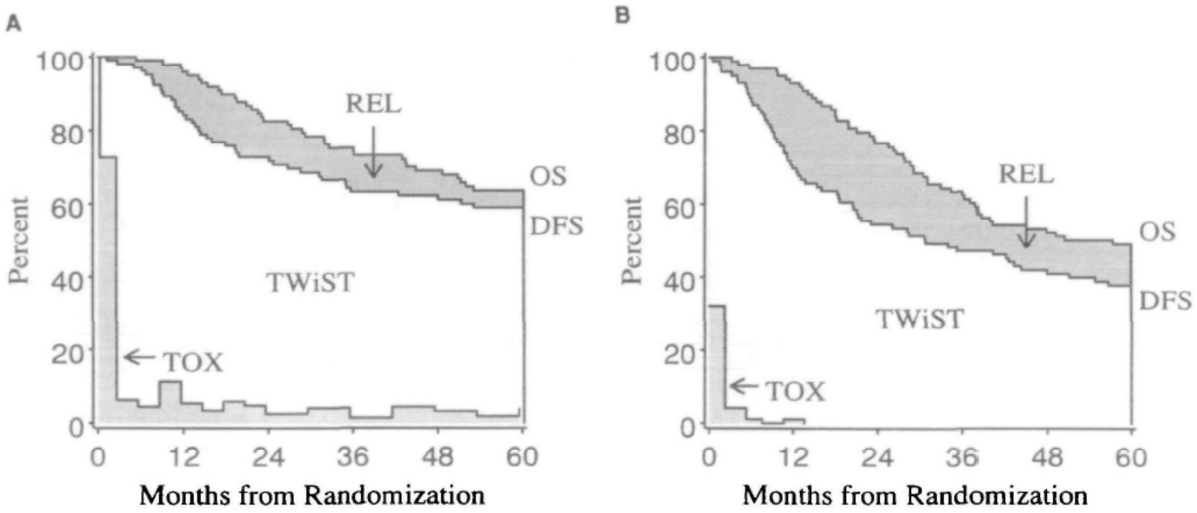


Figure 1: Survival plots for patients with resectable rectal cancer treated with, a) adjuvant chemotherapy and radiation therapy or b) adjuvant radiation therapy alone. Overall survival (OS), time with toxicity (TOX), disease-free survival (DFS), time without symptoms or toxicity (TWiST), and time after relapse (REL).
Source: Gelber RD, Goldhirsch A, Cole BF, Wieand HS, Schroeder G, Krook JE. A quality-adjusted time without symptoms or toxicity (Q-TWiST) analysis of adjuvant radiation therapy and chemotherapy for resectable rectal cancer. *J Natl Cancer Inst*. 1996 Aug 7;88(15):1039-45.

There is no doubt that Q-TWIST is a comprehensive method, where the weighting of different physical states enables a useful form of a benefit-risk comparison. The estimation of utilities is subjective and can therefore vary between different stakeholders. The method could potentially be extended to include more criteria. Also, the method has the potential to be further developed to clinical fields other than oncology. However, Q-TWIST does not incorporate uncertainty in the calculations and the model relies on point estimates. As with all models/methods/tools that calculate a single number, information is lost on the way.

Although Q-TWIST and QALY (Quality-Adjusted Life Year) differ in terminology, both methods attempt to make some kind of overall estimate of mortality and morbidity of HRQL (health-related quality of life)[58]. QALY is more broadly applied, while Q-TWIST is more or less restricted to oncology.

A year of perfect health is defined as 1 QALY, while immediate death is defined as 0 QALY. Two years of 0.5 QALYs are comparable to one year of 1 QALY. States worse than death can be defined and they would have a negative value, which will be subtracted from the overall number of QALYs. QALYs are frequently used by different institutions, e.g. NICE

(National Institute for Health and Clinical Excellence) in the United Kingdom (UK), in cost-effectiveness and cost-benefit analysis. QALYs are flexible and can be used in conjunction with other methods. Thompson and co-workers, for instance, used QALYs in conjunction with Markov models in a risk-benefit analysis of natalizumab[10;50-53].

A QALY in itself is insufficient in a complex decision-making process and is weak in incorporating uncertainty into the decision process. The definition of health states can differ between stakeholders, and QALYs are often criticised for being crude in their definitions. However, they can be used in conjunction with other methods. QALYs are especially well known and are used in Health Technology Assessments (HTA).

In conclusion, all these measures represent a need for incorporation of some kind of value judgement in the assessment of data. However, as discussed earlier, there is no real value in a benefit-risk setting in calculating one single number.

2.2.3 Survey methods

Randomised clinical trials (RCT) often ask few, specific questions about the efficacy of a new treatment. However, from a public point of view, it is more important to know whether the treatment is better than an existing one. Comparative Effectiveness Research (CER) tries to answer that question and it focuses on effectiveness, rather than efficacy.

CER is described [59]: *"CER is the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat and monitor a clinical condition, or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and populations levels."*

There is no straightforward way of performing CER. Several methods already exist and new methods are constantly emerging[60]. To mention one method, the Practice-Based Evidence Study Design (PBE-CPI) incorporates natural variation within data from clinical

practice to determine what works, for whom, etc. PBE-CPI is comparable to observational studies and consists of 7 steps[61]:

1. All interventions are taken into account to estimate the relative contribution of each
2. There are no specific hypotheses, but more general ones
3. Very broad inclusion criteria exist
4. Patients are characterised in a detailed way in relation to co-morbidity and functional status
5. Patient differences are controlled statistically, rather than through randomisation
6. Trans-disciplinary; involving different professions
7. Transparency for all stakeholders

CER and other methods for benefit-risk assessment of medicines should not be seen as diametrically opposite, but more in conjunction with each other. Each method has its rightful place during drug development and a drug life cycle[51].

There is an increasing indirect focus on effectiveness from regulatory agencies, but the current approval phase is very strictly controlled and leaves little room for CER. The idea of looking at a treatment effect on individual levels, instead of averages and mean values with standard deviation, has a new and interesting appeal.

RCTs are used both in drug development and in the post-market phase. CER could replace the use of RCTs in the post-market phase and could create a more dynamic picture of the effectiveness of drugs. However, in my opinion, the use of CER in drug development is all too embracing and might halt innovation and drug development. It is important to bear in mind that a new drug with a slightly better efficacy can seem to have little value, but it may become the ancestor of an entire new generation of innovative treatments in the long run.

Conjoint analysis (CA) resembles CER and originates from the business area of marketing, but has found its way to the health sector over the last couple of years. It was originally developed and used in market research to map the customer preferences. The

method is gaining widespread use in health care. Patients are usually not involved in the medical decision-making process, but this method unveils patients' preferences and confirms the relative importance of different aspects of a given treatment, how people weigh different aspects of a treatment, and it calculates overall scores for different treatment alternatives, etc. The direct use of CA in drug development is limited. However, it could be useful in illuminating patient preferences. Such preferences could then be used in the development of drugs. Indirectly, CA could thus have an impact on drug development[10;62;63].

The direct use of CA and CER in a benefit-risk assessment setting is limited, but the ideas of mapping different stakeholders' preferences and incorporating the results into the benefit-risk analysis seem to be self-evident aspects.

2.2.4 Early benefit-risk methods

In 1996 Edwards and co-authors[64] published a paper on a simple benefit-risk assessment method, which they named a "merit assessment". The motivation for the development of this method was due to the lack of transparent assessment of risks and benefits from clinical trials as well as the lack of a structured way to handle AEs related to post-marketed drugs. The goal was to create a simple method that allowed general ideas of the possible outcomes of a given drug to be conveyed, resulting in a comparison of medicines. Secondly, the aim of the method was to pinpoint areas, where data should be collected post-marketing. This should allow a dynamic benefit-risk assessment that is upgraded as data from the use of the medicine is updated.

The method consists of a so-called "principle of threes" based upon the concepts of seriousness, duration and incidence related to the benefits of the drug and the risks associated with taking it. Each criterion is weighted using a simple scale of high, medium and low, as seen in Table 1.

Table 1: The qualitative grading system used in the "Principle of threes"

Gradation	High	Medium	Low
Seriousness	Fatal	Disabling	Inconvenient
Duration	Permanent	Persistent	Temporary
Incidence	Common	Frequent	Rare

Source: Edwards R, Wiholm BE, Martinez C. Concepts in risk-benefit assessment. A simple merit analysis of a medicine? *Drug Saf*. 1996 Jul;15(1):1-7.

The three most important benefits and risks are selected based on the context of the disease or indication. The scoring is subjective and scores are assigned based on the same scale used for weighting.

Table 2: An example of how the grading system from Table 1 can be used

Gradation	High	Medium	Low
Disease			
Seriousness		X	
Duration	X		
Incidence			X
Level of improvement produced by drug			
Seriousness	X		
Duration		X	
Incidence			X
Adverse effects of the drug			
Seriousness			X
Duration			X
Incidence		X	

Source: Edwards R, Wiholm BE, Martinez C. Concepts in risk-benefit assessment. A simple merit analysis of a medicine? *Drug Saf*. 1996 Jul;15(1):1-7.

In the report by CIOMS working group IV from 1998, the Transparent Uniform Risk/Benefit Overview (TURBO) is mentioned. No published scientific papers have been found on this

method or its use. The report refers to personal communication with Dr Willem Amery[1]. As with the “Principle of Threes”, this method is motivated by a frustration over the lack of methods for a more structured approach to benefit-risk assessment.

The TURBO method assigns scores to the most serious adverse event and one additional risk (R-score) as well as the primary benefit and one ancillary benefit (B-score). The scoring scale is 1-5 for the primary benefit/risk and 0-2 for the secondary benefit/risk. The weighting is relative and no scale is mentioned. The results are plotted in a XY-diagram, as seen in Figure 2, and a “T-score” (therapeutic score) can be determined from the diagram.

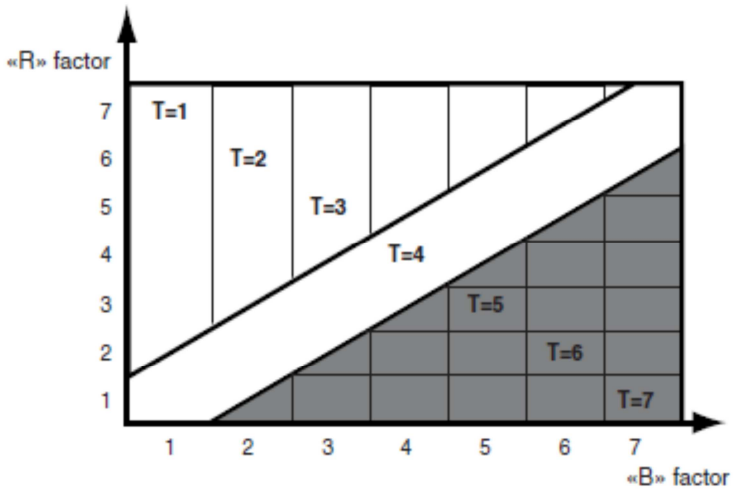


Figure 2: The so-called “TURBO-diagram”. The overall score, T-score, can be determined from the diagram.
Source: The Council for International Organizations of Medical Sciences (CIOMS). Working CIOMS Group IV - Benefit-Risk Balance for Marketed Drugs: Evaluating Safety Signals, 1998.

The method is very simple and is a transparent approach to assessing the most important benefits and risks. However, in most cases it is very difficult to reduce a benefit-risk assessment to a few criteria. The scoring is subjective and there are no guidelines on how these should be assigned. Without additional elements, the method is not suitable for comprehensive benefit-risk assessment.

Both methods, “Principle of Threes” and TURBO, are sound attempts to structure a benefit-risk assessment and the scales are coarse, reflecting the uncertainty related to data.

However, there are several drawbacks: first of all, both methods are restrictive in a sense that one can only include very few benefits and risk criteria; secondly, there is no elaboration on how the scores are assigned and what defines a clinical relevant difference between two treatments [1;5]. The two methods have neither been generally accepted nor used.

Experience from the approaches discussed so far clearly shows that a decision analytic methodology must be used in a benefit-risk assessment. A more comprehensive approach where multiple considerations are made against each other is much needed.

2.2.5 Comprehensive approaches

Over the last decade, several comprehensive methods have been developed. They all seem to find inspiration in Multi-Criteria Decision Analysis (MCDA), which is deeply rooted in Decision Analysis (DA). Ronald A. Howard defined the modern version of decision analysis in 1964[65]. The fundamentals of decision analysis, on the other hand, go hundreds of years back and have connections to Daniel Bernoulli's (1700-1782) probability and statistical theory, Bayes' theorem by Thomas Bayes (1701-1761) and Pierre-Simon Laplace (1749-1827), statistical decision thinking by Abraham Wald (1902-1950), and many more. Laplace apparently also developed Bayes' theorem without knowing of its existence.

It is out of the scope of this thesis to evaluate and discuss decision analysis and its history. Instead, the focus will be on modern developments, which will be divided into two categories: 1) Multi-Criteria Decision Analysis (MCDA) and methods inspired by it, and 2) approaches that are inspired by and related to decision analysis.

2.2.5.1 Multi-criteria decision analysis

MCDA has, as mentioned, its origins in decision theory. Keeney and Raiffa published a book on MCDA in 1976[20] and are often considered as the founders of this discipline. Keeney later published another book, *"Value-Focused Thinking: A Path to Creative Decisionmaking"*[18], in which he focuses much more on the aims and objectives, instead of decision alternatives,

etc. In 2002 Belton published a book simply entitled *"Multi-Criteria Decision Analysis: An Integrated Approach"*[17], in which different versions of MCDA are evaluated and discussed.

MCDA breaks down complex problems, analyses and weighs each piece individually, puts the pieces back together and displays the result for the decision makers. There exists several different variants of MDCA[17;66] and MCDA is widely used in economic and political decision making[19;21]. Several methods in the literature are related to or inspired by MCDA, e.g. Clinical Utility Index (CUI)[67-71], Desirability Utility Index (DUI)[72], Global Benefit:Risk Score (GBR score)[73;74] and Benefit-Less-Risk (BLR)[10]. It is needless to evaluate all these related methods; the focus will instead be on MCDA.

Over the last couple of years, a specific version of MCDA, the MCDA Value Tree, has gained interest from several different stakeholders within medical decision making[9;10;12;13;16;23;23;75-79]. MCDA Value Tree uses an algorithm that combines weighting and scoring of specified criteria. The method consists of eight successive steps, as described in the following[19]:

1. **Establish the decision context:** It is very important that the purpose of the assessment is clearly defined. Otherwise, there is the risk of performing a perfect analysis for the wrong problem. However, the decision context is dynamic and can be changed throughout the assessment as new issues emerge, which may indicate a change of the aim. A clear aim helps keep the assessment on track. The key players/stakeholders to take part in the analysis are identified and they represent anyone who can make a useful and significant contribution to the assessment.

2. **Identify the options to be appraised:** Different options within the boundaries of the decision context are identified. They could be, e.g. different drugs for the same indication.

3. **Identify objectives and criteria:** Criteria are defined as the consequences of options. Strictly speaking, it is the criteria that are assessed and not the options themselves. Criteria are specific measurable entities, measured either subjectively or objectively, expressing the multiple ways that an option can create value. All criteria are organised by clustering them under higher-level and lower-level objectives in a hierarchy (value tree), as seen in Figure 3. When objectives and criteria are organised in a value tree, conflicts amongst objectives are often seen, which leads to an iterative process of refinement of the value tree.

4. **Scoring:** The first step in comparing different criteria is to assign a numerical value to each criterion based on its performance. This is done by constructing scoring scales, e.g. preference scales, which are simple scales that are anchored at their ends by the most and least preferred value. The scale can, e.g. be from 0-1 or 0-100. Scales are constructed for each type of criterion.

5. **Weighting:** The second step in the comparison of different criteria is to assign weights to each criterion and the most used method is “swing weighting”. Weights are used to reflect the relative preference for a criterion. This is based on comparisons of differences. To enable a comparison, the participants in the assessment are forced to take into account both the difference between the least and most preferred options and how much they care about that difference. It is important to remember the difference between measured performance and the value or weight of that performance in a specific context. In other words, a given large difference in performance for a criterion between two options may not have any added value at all. In a value tree, with multiple levels, relative and cumulative weights must be defined. Relative weights are assigned within families of criteria that originate from the same parent criterion. The cumulative weight of a criterion is the product of its relative weight and the relative weight of its parent.

6. **Weighted scores:** Overall weighted scores, *S*, for each option can be calculated by multiplying scores and weights and adding them together:

$$S = \sum_i w_i s_i$$

However, it is very important to note that the simple weighted scores calculation shown above is justified, only when all criteria are mutually preference independent. This is straightforward. If several criteria are correlated, the overall weighted score is biased, since some criteria will be counted multiple times.

7. **Results:** The results are shown in a value tree, and overall weighted scores of each criterion for each option can be calculated and shown. A benefit:risk ratio can also be calculated for each option.

8. **Sensitivity analysis:** The purpose of a sensitivity analysis is to provide evidence for the actual results. A sensitivity analysis should reveal whether or not the results are consistent with respect to any change in the model. Weights can be changed in the whole range of the scale to investigate any impact on the overall results. The impact of missing information should also be investigated. Any uncertainty and its impact on the results should clearly be analysed.

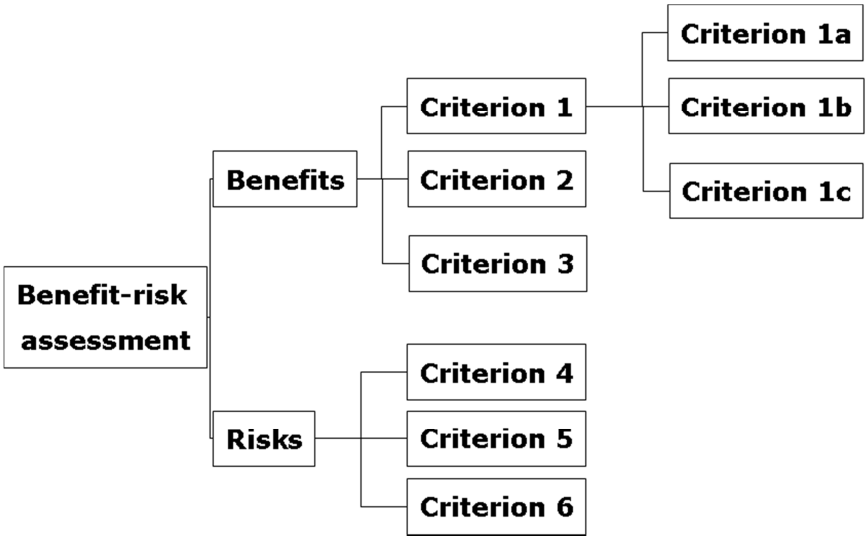


Figure 3: An example of a value tree for the benefit-risk assessment of medicine

MCDA can combine numerical judgements and data, visualising the results in a value tree or as a benefit:risk ratio. The method can also accommodate several criteria and can be very comprehensive. It is flexible, since criteria can be added, moved or deleted. It encourages discussions on what is important to consider in a given assessment. Furthermore, the discussions about the subjective weights and scales give indications about which aspects of the drug are important to focus on.

However, there are several concerning features of this method when used in medical decision making, as expressed by CHMP[9]. First of all, the end product is a ratio, which in reality can be very difficult to interpret, because data is added, multiplied, divided, etc. every time information is lost. Furthermore, the method performs linear combinations of single values, which can be very misleading in very dynamic and non-linear systems, like the human body. It is assumed that all criteria are mutually independent, which is generally an incorrect assumption when dealing with biological systems, where most processes are more or less interconnected.

Including several criteria, which are coupled, may induce unintended effects on the weighted scores. Merging criteria to avoid correlations would solve this problem. However,

this complicates direct measurements of criteria variables, e.g. from clinical trials. Correlations are difficult to account for in an MCDA value tree.

The tree structure complicates the placement of criteria and restricts the decision model. The tree structure itself imposes several problems. Depending on how one structures the chosen criteria in the value tree, different results can be obtained. Furthermore, by including more criteria in the value tree, the impact of the single criterion on the end result diminishes with the increasing number of criteria, which means that a criterion that stands out from the crowd simply vanishes in the end result. Of the bottom level criteria, one criterion will count as:

$$\frac{w_x s_x}{w_n s_n + w_y s_y + ... + w_z s_z}$$

This fraction will only be large if the score and all weights leading to the given criterion are large. If the weight for one criterion is large, all other criteria will consequently have smaller weights. To include all clinical data, criteria must be explicit, e.g. SAEs should probably consist of several sub-branches. But as more criteria are included in the value tree, they will be more marginalised. This may be acceptable for economic and political systems, but for biological systems it is unacceptable, since one single criterion may and should have the ability to decide the fate of the drug.

The marginalisation problem can be overcome by using a score range with a large or no upper limit, e.g. exponential or asymptotic functions. However, it may be insurmountable to construct such functions for a great number of criteria. Another approach would be to assign the weights based on the scores. However, this limits the usability of weights, since we would effectively be “cheering for the winner”.

In the value tree, the placement of criteria is of crucial importance. Low weights on top level criteria lead to reduced weights of lower levels. This is especially important if a criterion can be placed under different branches, since the results will depend on the structure of the

tree. Introducing extra criteria will reduce weights for other criteria, because the weights are normalised.

Furthermore, it is practically difficult to handle multiple trials, since the value tree will become too complex, and the impact of correlations will be difficult to control. Therefore, individual value trees must be constructed for each trial; when a new drug is submitted to an agency, the submission dossier usually consists of several trials. A more comprehensive approach is much needed. It would therefore enhance the overall view if all the results could be combined in an overall assessment.

The scales used in MCDA are subjective and impose several problems. If they are created before data are revealed, they tend to be large and results gather in the middle and cannot be distinguished. If they are created after data is revealed, they have a tendency to be narrow and there will always be a difference between two drugs[80].

For each criterion, a value function must be defined that describes the importance and desirability of achieving different performance levels[17;18;80]. A value function describes the best and worst scenario and often it is assumed that there is a linear correlation between these two options, as seen in Figure 4.

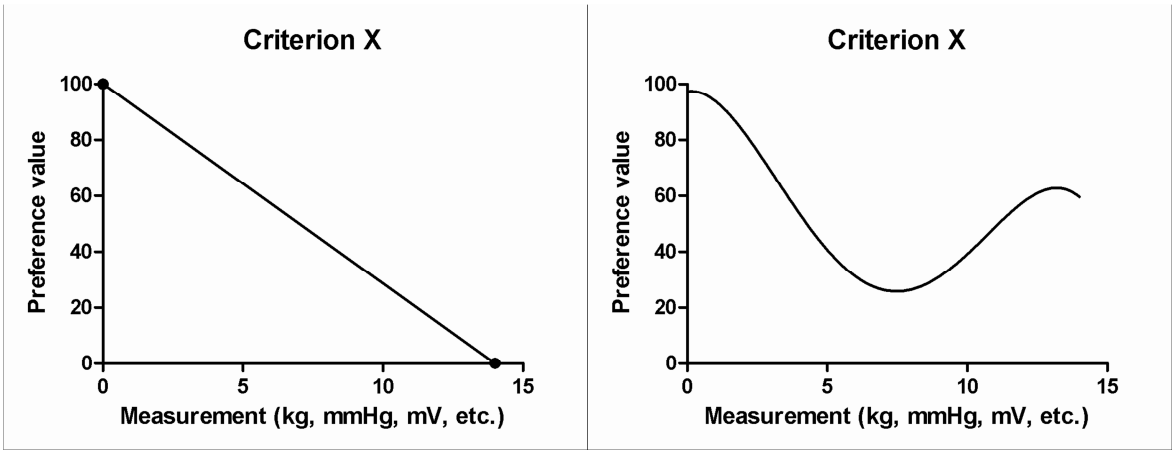


Figure 4: a) The value function for a criterion with linear relations, and **b)** the value function for a criterion with non-linear interconnections.

The nature of these value functions is not always self-evident for living systems. This clearly demonstrates the fact that MCDA was originally developed for political and economic decision making and it would be naïve to believe that a swift transition to medical decision

making can be made without problems or issues. Further research is needed to reveal the nature of different value functions for different criteria in dynamic biological systems.

Uncertainty related to data is not incorporated in MCDA. Weights and scales for scoring are subjective and can vary from stakeholder to stakeholder, compromising the true value of the results of an assessment. When a sensitivity analysis is performed, it is the variability of the weights and the impact it has on the results that are investigated and not the uncertainty associated with the data itself. Again, this clearly demonstrates that MCDA in its present form is intended for economic and political decision making and is less suited for biological systems.

However, despite these caveats, the method has a structured approach to benefit-risk assessment, which will undoubtedly increase transparency compared to a non-structured approach. The structured way to approach a problem/decision/assessment will later prove to be very useful in clinical benefit-risk assessment of drugs and therapies. MCDA has given inspiration to many current initiatives around the world that will be discussed later[16;81;82].

2.2.5.2 Decision trees, Markov models and Bayesian theory

Most decisions affect other alternatives and consequences in other decisions. In a specific decision, it will not be justifiable to include every possible scenario. A single decision tree is created for a specific decision. From a clinical/medical point of view, these decision trees exist in the mind of every doctor. In emergency rooms (ER), so-called flow charts can be found to aid snap decision making under pressure when treating acutely ill patients.

The difference between flow charts and decision trees is that decision trees include probabilities, where each choice is linked to a probability, e.g. in Bayesian decision theory/trees. During the last 25 years, Bayesian statistics have moved into almost every corner of medical sciences. Bayesian statistics have been adapted from simple clinical trials to meta-analysis, from survival modelling to molecular genetics, etc.[83] The use of Bayesian

statistics in decision making has clearly been demonstrated in the literature[84]. However, the practical use of Bayesian statistics is still an issue of discussion and the regulatory agencies have so far been reluctant to use Bayesian statistics on a large scale. The main argument has been that Bayesian statistics are premature and cannot cope with modern-day requirements. However, the enormous development within this field and the obvious fragile aspects of classical statistics may lead to a shift of paradigms.

The essence of Bayesian statistics and decision theory is the definition of prior beliefs/probabilities and the calculation of posterior beliefs/probabilities based on these. In Bayesian decision making, the following are identified[84]:

1. The decision maker(s)
2. The possible actions
3. The uncertain consequences
4. The possible sources of evidence
5. The kinds of utility assessment required

In Figure 5, an example of the breast cancer diagnosis process is shown. When a decision is made to screen for breast cancer, the outcome is either positive or negative. There are probabilities associated with these outcomes and so forth[18].

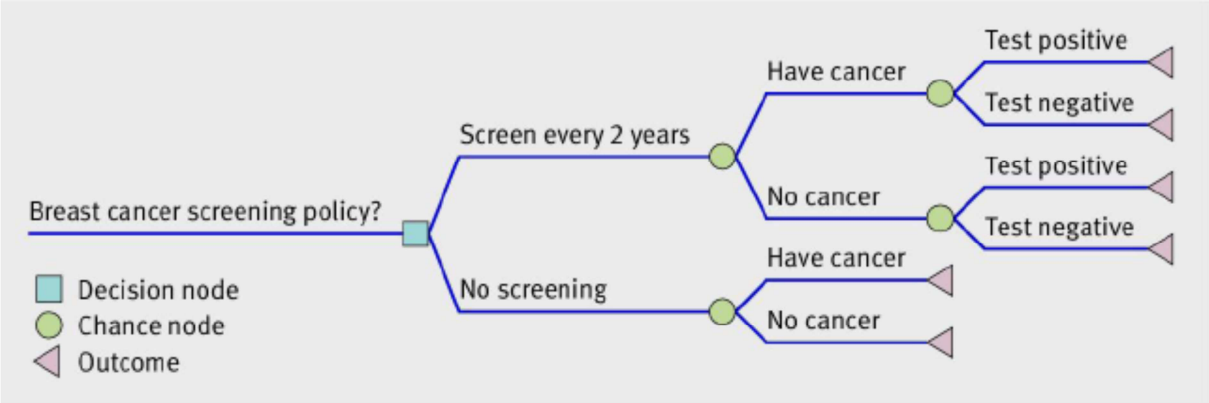


Figure 5: Decision tree for breast cancer showing the decision nodes, chance nodes, outcomes and the probabilities of each level.
Source: Petrou S, Gray A. Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting. *BMJ*. 2011;342:d1766

Markov models resemble decision trees, where health states are defined and incorporated in a tree-like structure, as described in the following. Markov models assume that a patient is always in a finite number of health states, which are clearly defined. A patient can experience a given event more than once. The risk of an event is persistent over time. Each health state is associated with a given probability. A simple tree of the consequences of anticoagulation treatment is shown in Figure 6, where three health states exist. Only certain transitions can occur and a patient can move from the well state to the disabled state and back again.

The Markov model consists of several cycles and the length of the cycle is chosen to resemble a clinically relevant time length. The model conducts multiple cycles until the patient is in the death state, from which there is no return. In the Markov model terminology, such states are called *absorbing states*. Different utilities can be assigned to each state and based on the different amounts of time the patients spend in each health state, the amount of quality-adjusted life expectancy of the patient can be calculated.

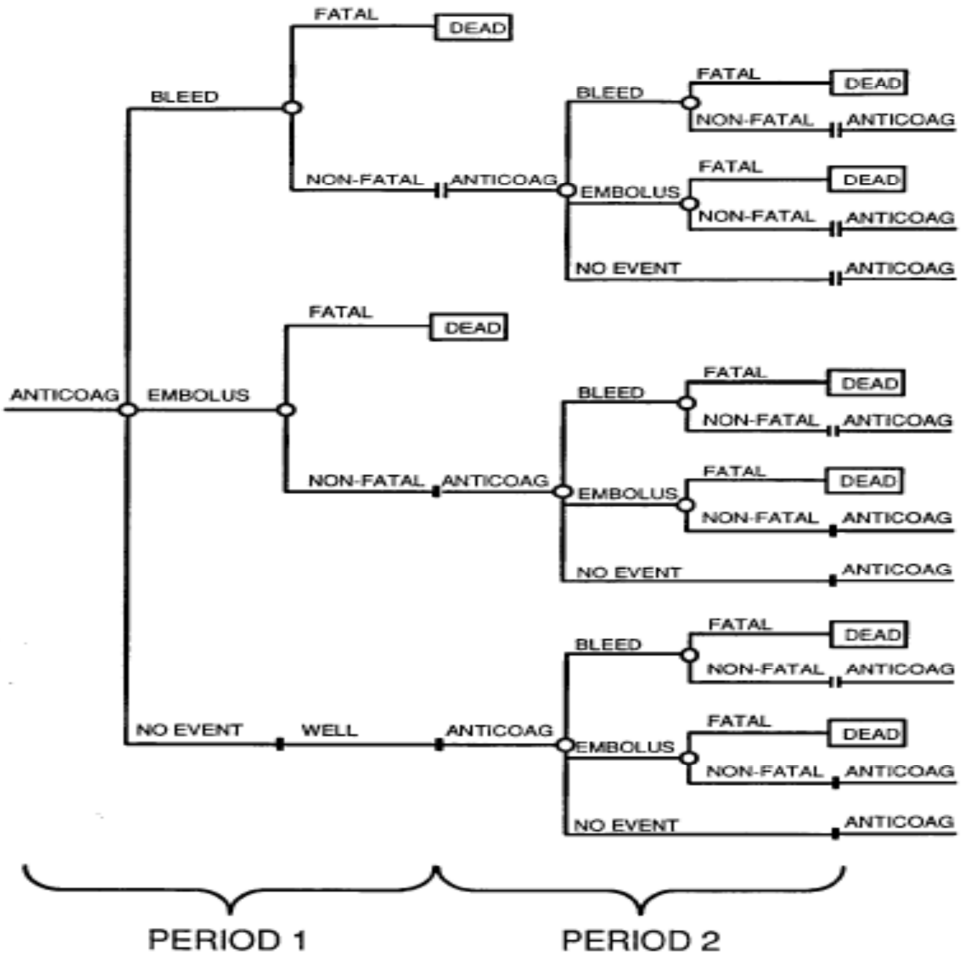


Figure 6 : A simple tree of the consequences of anticoagulation treatment is shown.
Source: Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. *Med Decis Making*. 1993 Oct;13(4):322-38.

An event that has short-term effects and has transitions to other states, but not itself, is called a *temporary state*. After each cycle, the probability of each state may change depending on the previous events. The Markov model also assumes that during a single cycle, each patient undergoes only one state transition[53;85;86].

Decision trees are excellent for single decisions, e.g. should we approve this drug now or wait three more months for additional data?, is drug A better than drug B for the primary endpoint?, etc. For complex decisions, more complex decision trees are needed and this requires expertise and time. However, in a benefit-risk assessment, we are not looking for a go/no-go decision, but are trying to gather all the pieces of the puzzle and put them together

to support the intellectual decision-making process. It is the art of enabling the human mind to comprehend an enormous amount of data.

In Bayesian decision theory under step 3, statements regarding prior probabilities are made. These can be made based on prior knowledge about the disease or risk of a condition. However, if there is a lack or limited amount of evidence, the statements about the prior probabilities can be questioned and the use of the method is restricted. Bayesian decision making can handle a single criterion at the time and does not handle multiple criteria, but issues such as identification of uncertainties, decision makers and possible sources of evidence are worth remembering.

Markov models depend on clearly defined health states that are associated with different probabilities. To make any assumption of these probabilities, there is a need for substantial background information/data. The use of Markov models in drug development may therefore be restricted. However, in the post-market phase of a drug's life cycle, where more data are available, the use of Markov models may be justified.

2.3 Discussion and Conclusion of Chapter 2

The most obvious tools/methods/models from a benefit-risk assessment point of view have been divided into main categories and evaluated. The aim was to investigate not just what the current methods are missing, but also which aspects of existing methods/tools/models we can use to enhance 1) transparency, 2) objectivity and 3) communication of results and decisions in our future benefit-risk assessment method.

None of the evaluated methods focused on how to capture clinical significance and relevance of data. As discussed earlier in this thesis, while statistical significance is important, clinical relevance of data is crucial[25].

With regard to metrics, they are very simplistic and it is unsuitable to calculate one overall value for a drug. From a benefit-risk assessment point of view, there is no value in

calculating a single value, e.g. a benefit-risk ratio. Too much information is lost in the process. Several of these problems are also observed in quality measures, where single values are calculated.

Survey methods focus on, e.g. patient preferences, alternative treatments, etc. While all these issues are obviously relevant and important, there is a lack of bringing them together in a structured way to enable a transparent benefit-risk assessment. Some early approaches, TURBO and Principle of Threes, attempt to cope with this. The attempts are acknowledged, but unfortunately both approaches are simplistic and fail to capture clinical significance and provide transparent visualisation of results. Furthermore, little attention is paid to how to handle uncertainty.

The more comprehensive approaches such as MCDA, GBR and BLR try to cope with multiple criteria, but nonetheless end with a single value or ratio. In, e.g. MCDA, little attention is paid to how to cope with uncertainty. In the MCDA Value Tree, which has inspired many current developments, the correlation between criteria, the problems with marginalisation and value functions, and the lack of visualisation make it difficult to adapt to complex biological issues.

Looking beyond the critique, several aspects from different methods reveal useful potential in a benefit-risk assessment. The incorporation of weighting is a key element in the direct comparison of benefits and risks. The definition of a decision frame is pivotal to avoid confusion in the decision process. An evaluation of uncertainty and evidence, where all the weaknesses and strengths are communicated in transparent ways, will give credit to any given conclusion one might reach in the assessment.

A benefit-risk assessment should not merely be a summary of the main statistical findings, with differences in mean averages being the main focus. Instead, the assessment should be a complement to the statistical analysis; a necessary focus on clinical significance and the individual subjects is therefore needed. Simplicity is to favour, because it increases transparency and communication. A complex method/model often gives a false impression of accuracy with regard to the conclusions reached.

Whatever weaknesses MCDA and other related methods might have, the structured process in itself increases transparency in the decision-making process.

Current assessments of interventions seem to be inadequately dealt with; there is a lack of justification of choices, minimal transparency of the process and the end results and limited focus on clinical significance. A discussion of clinical significance is important to include in an assessment, since it increases both transparency and communication of decisions. None of the evaluated methods fulfil all pivotal elements to produce a consistent, structured and transparent benefit-risk assessment with focus on clinical significance of data and visualisation of decisions and results of the assessment.

3 Current initiatives

There are international guidelines approved by regulatory agencies, e.g. International Conference on Harmonisation (ICH) guidelines[87], which clearly state what is required from the industry in order to gain approval of a drug. Nevertheless, regulatory agencies differ in their decisions on the same dataset. It is occasionally seen that one agency approves and another disapproves a new drug candidate.

In recent years, there has been some product withdrawals due to serious adverse drug reactions[88]. Consequently, a debate has started, both internally in regulatory agencies and externally in the general public and in the pharmaceutical industry, questioning the quality, consistency and robustness of the benefit-risk assessments made in relation to the approval of drugs.

The field of benefit-risk assessment is also gaining increased attention from pharmaceutical companies. The pharmaceutical industry invests millions of USD in drug development, and thousands of patients are enrolled in clinical programmes even though the failure rate is extremely high. First of all, from an economical point of view, a new method that can predict or minimise the risk and that can reduce the failure rate during clinical drug development will have an enormous commercial impact. Secondly, but equally important, from an ethical point of view, a new method could reduce the unnecessary enrolment of patients into clinical trials that have a high failure rate. A new method can hereby improve patient safety in clinical trials.

This has led to several initiatives at regulatory agencies and in the industry and there are probably even more on-going initiatives that are not published or otherwise communicated to the public. Some of the main initiatives have been taken by EMA, CIRS, the Pharmaceutical Research and Manufacturers of America (PhRMA), and IMI (Innovative Medicines Initiative). Other initiatives are based in FDA, ISPOR, and the consortium consisting of Health Canada, Swissmedic, TGA and HSA (the COBRA consortium).

3.1 FDA

Several documents describing the current recommendations and guidance were identified, as described in Chapter 2. These documents cover pre- and post-marketing phases of drug development and approval. There is no explicit guidance on how benefit-risk assessment should be conducted. Therefore, the assessment is qualitative and influenced by, e.g. personal expertise and experience.

This has led to an increased internal awareness to improve the current status of benefit-risk assessment. FDA currently has on-going activities that aim at developing a qualitative structured approach to the benefit-risk assessment of drugs[10]. This approach is grounded in the qualitative analysis of data. In Table 3 the preliminary work is shown. Key issues related to the drug of interest are identified and discussed. The aim is to gain an overall view of the benefit-risk balance of a given drug.

The framework is simple and addresses critical issues in a benefit-risk assessment. It can improve communication both internally and externally. However, the expert judgements made are purely qualitative and might not always be transparent. The framework is still under development and it is premature to make any final statements about its pros and cons.

Table 3: FDA's preliminary framework for the benefit-risk assessment of drugs. The framework is meant as a tool for regulators.

Consideration	**Favourable benefit-risk**	**Non-contributory**	**Unfavourable benefit-risk**
Severity of condition			
Unmet medical need			
Clinical benefit			
Risk			
Risk management			

Source: Benefit-risk methodology project - work package 2 report: Applicability of current tools and processes for regulatory benefit-risk assessment. EMA/549682/2010.

3.2 EMA

A report in 2008 from the CHMP[9], reviewing existing methods, concluded that the benefit-risk assessment of new drugs is a complex process that requires evaluation of a large amount of data. The report analysed existing methods such as Principle of Threes, TURBO and MCDA for the benefit-risk assessment of drugs. The methods were tested for critical aspects, e.g. consistency, comprehensiveness, transparency and communication of the results of the assessments made by the CHMP.

Furthermore, it concluded that expert judgement will still be the cornerstone of benefit-risk evaluation for the authorisation of medicinal products. A qualitative framework with a structured process is discussed, where the most important benefits and risk are identified more clearly and where the importance of these benefits and risks is explicitly described. It is uncertainty evaluated and is quantified in the best-case scenario. MCDA is mentioned as a possible and promising way to cope with multiple objectives and criteria. Nevertheless, CHMP express concerns, such as the linear combinations of single values and the lack of attention to uncertainty, and conclude that the method focuses on decision making, where overall values are compared, rather than looking at the elements individually. Based on their review of some of the existing methods, CHMP thus concludes that there is a need for a tailored method for medical benefit-risk assessment in relation to drugs and that further research within the field of benefit-risk assessment is needed and that experts and assessors should be involved.

This has led to the initiation of the Benefit Risk Methodology Project in 2009. The project consists of five work packages and is expected to be completed by the end of 2011. The five work packages are as follows:

1. Current practice of benefit-risk assessment for centralised procedures in the EU regulatory network.
2. Applicability of currently available tools and processes for regulatory benefit-risk assessment.

3. Adaptation and field testing of recognized tools and processes (from WP 2).
4. Development of a benefit-risk tool/method that can add value in the regulatory process.
5. Development of a training package on the new tool/method for regulatory assessors.

To date, the first three packages have been completed. The first work package[34] was concerned with how benefit-risk assessment was being defined and conducted in the participating agencies. Some of the conclusions were that there are no common definitions of terms such as benefit and risk, that no formal guidance on the process of their balancing exists and that the formulation of the benefit-risk balance was intuitive, based on personal experience and expertise. The project team developed a qualitative model for the assessment of uncertainty, as seen in Table 4. This model has been implemented in the Day 80 Assessment Report (D80 AR). However, how to perform an evaluation of uncertainty is not described in this report.

Table 4: The four-fold qualitative model for the assessment of uncertainty.

Favourable effects	Uncertainty of favourable effects
Unfavourable effects	Uncertainty of unfavourable effects

Source: Benefit Risk Methodology Project - work package 1: Description of the current practice of benefit-risk assessment for centralised procedure products in the EU regulatory network. EMA/213482/2010.

The second work package[10] was an evaluation of existing methods/models for their applicability for regulatory benefit-risk assessment processes. A report was published in the second half of 2010. The included methods were evaluated based on a set of predefined criteria: 1) logical soundness, 2) comprehensiveness, 3) acceptability of results, 4) practicality, and 5) generativeness. The report concludes that any quantitative method requires a qualitative framework. Only three approaches were identified, which were sufficiently comprehensive enough: Bayesian statistics/Bayesian decision theory, decision trees and influence/relevance diagrams, and MCDA. The MCDA was seen as the most

promising method. Furthermore, five other approaches with a more restricted scope were also identified.

The third work package[89] was published recently and it describes the field testing of MCDA, Decision trees, Markov models, and probabilistic simulation. The authors conclude that a qualitative framework, called the PrOACT-URL, should be adopted by the agency and that more comprehensive methods, such as MCDA, could be used during a CHMP meeting. Furthermore, the pharmaceutical industry is to be encouraged to submit quantitative benefit-risk assessments in their applications. In the second report by EMA's benefit-risk methodology project, the PrOACT-URL (**Pr**oblem, **O**bjectives, **A**lternatives, **C**onsequences, **T**rade-offs, **U**ncertainty, **R**isk and **L**inked Decisions) is described as an eight-step qualitative framework that provides the definition of what is meant by a comprehensive method[10].

The conclusions from the remaining work packages will reveal the final road that the project recommends.

3.3 Centre for Innovation in Regulatory Science (CIRS)

CIRS was established in 2002 and was called *CMR International Institute for Regulatory Science* until 2010. The centre is an independent corporation with its own dedicated management and advisory boards. Its funding is derived primarily from membership dues.

During its lifetime, CIRS has established a leadership within the field of benefit-risk assessment not only as a forum for discussions between regulators and the industry, but also in the development of a benefit-risk assessment method. Several workshops have been held addressing different aspects of benefit-risk assessment[24;75;77;90]. The first workshop was held in London (UK) in 2004 and was entitled ***"****Benefit-Risk Assessment: The Development of a Model for Benefit-Risk Assessment of Medicines Based on Multi-Criteria Decision Analysis"*. The centre has since held annual workshops focusing on different aspects of benefit-risk assessment each time. In recent years, CIRS has acknowledged that despite the

immense activity concerning the development of methodologies, there is a need for collaboration between major key opinion leaders.

Consequently, CIRS initiated an international collaboration and established the *"Task Force for the Exchange of Ideas on Initiatives for the Benefit-Risk Assessment of Medicines"* in 2011. All major regulatory agencies, such as FDA and EMA, and pharmaceutical companies with some kind of involvement in the development of benefit-risk methodologies have been invited and have joined the task force. Once again, CIRS has shown that it is the foundation whereupon different stakeholders meet, discuss and create consensus.

The centre also has an advisory role in relation to the COBRA group (**C**onsortium **O**n **B**enefit **R**isk **A**ssessment), and the method being tested by this group is based on a paper published by Liberti et al. in 2010[81]. The founder of CIRS, Prof. Stuart Walker, co-authored on the first text book on the topic of benefit-risk assessment: "*Benefit Risk Appraisal of Medicines: A Systematic Approach to Decision Making*"[5]. The centre has also published several other scientific papers[79;81;90]. The centre acknowledges the wide range of methods and opinions, and attempts to establish a forum and an environment, where a consensus can be made, as mentioned.

3.4 The COBRA consortium

A consortium, the COBRA (**C**onsortium **O**n **B**enefit **R**isk **A**ssessment) group, consisting of Health Canada, Therapeutics and Goods Administration (TGA, Australia), Health Services Authority (HAS, Singapore) and Swissmedic, has worked in close collaboration with Centre for Innovation in Regulatory Science (CIRS) since 2006 on the development of a common approach to benefit-risk assessment. The framework that is currently being tested is seen described in a paper by Liberti et al.[81], and is seen in Table 5. This framework is adopted and under further development and testing by the consortium.

The framework is grounded in the MCDA value tree. However, the COBRA group has not yet revealed the more precise details behind weighting, scoring, etc. It is therefore premature

to begin a discussion of its pros and cons, but any given structured approach to a problem will undeniably increase transparency in any given assessment.

Table 5: The framework developed by Liberti et al. and adopted by the COBRA consortium: a five-step approach.

Starting point	Data on efficacy and safety from company submission	
	Therapeutic indication	
	Options to be addressed	
Step 1	Construction of summary tables	
Step 2	Value tree	
	All possible benefits	All possible risks
Step 3	Assessment of importance and prioritisation	
	Selected benefits	Selected risks
Step 4	Assignment of values for each benefit and risk criteria for each option	
Step 5	Benefit-risk assessment	
	Expert judgement	

Source: Liberti L, McAuslane N, Walker S. Progress on the development of a benefit/risk framework for evaluating medicines. Regulatory Focus 2010 Mar;1-6.

3.5 Pharmaceutical Research and Manufacturers of America (PhRMA)

Since 2006 PhRMA has devoted attention to the development of a benefit-risk assessment framework. The Benefit-Risk Action Team (BRAT) was established with the aim of developing a framework that could be used both in the pharmaceutical industry and in regulatory agencies. They developed a six-step framework, as seen in Figure 7, which focuses on a qualitative assessment[10;15;82].

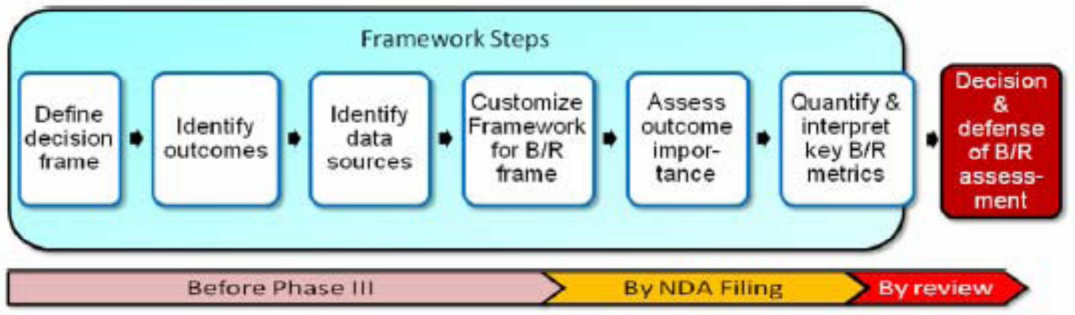


Figure 7: The BRAT framework developed by PhRMA.
Source: Benefit risk methodology project - work package 2 report: Applicability of current tools and processes for regulatory benefit-risk assessment. EMA/549682/2010.

The BRAT framework has some elements in common with MCDA and it uses a value tree that is converted to a key benefit-risk summary table. The first step is similar to the first step in MCDA, where the decision context is defined. The second step identifies benefit and risk criteria that are used to create the value tree in step 4. In step 3, data sources are identified and listed. In step 5, the relative importance of the different criteria is evaluated. Weights are assigned, but no clear method is provided. In step 6, the results are visualised in a so-called key-summary table, as seen in Figure 8.

The framework is advocated as a flexible framework with many advantages. There is no real use of a value tree; instead, all wanted criteria are transformed to the so-called “key benefit-risk summary” table. This is a positive aspect of the BRAT framework, since correlations are avoided in this manner. Criteria are organised in a hierarchical order, and one is referred to existing methods for weighting. The BRAT framework does not exclude any formal weighting.

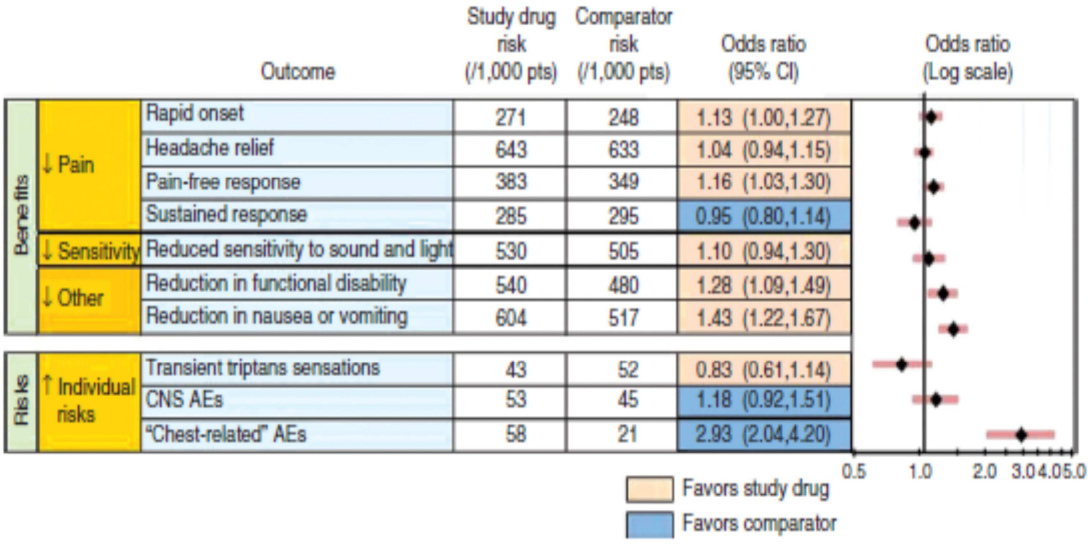


Figure 8: Key benefit-risk summary table in the BRAT framework with odds ratio.
Source: Levitan BS, Andrews EB, Gilsenan A, Ferguson J, Noel RA, Coplan PM, et al. Application of the BRAT Framework to Case Studies: Observations and Insights. Clin Pharmacol Ther 2011 Feb;89(2):217-24.

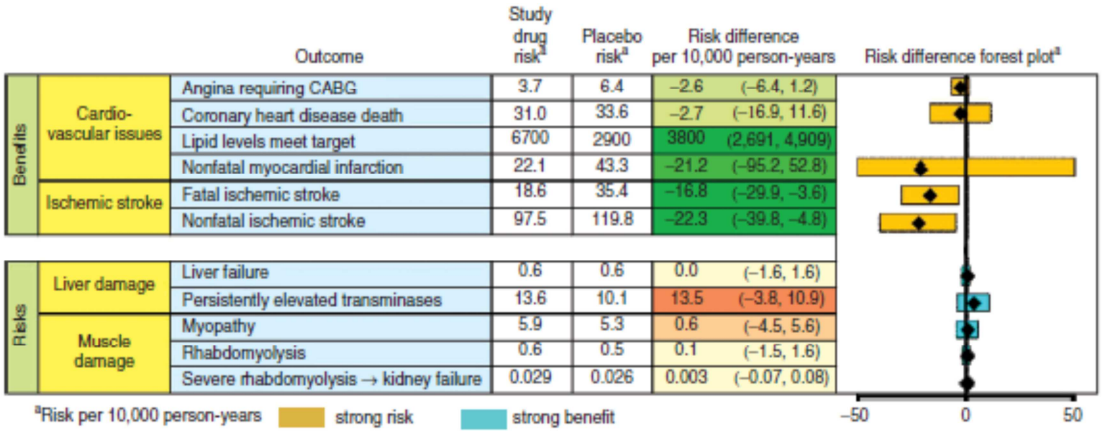


Figure 9: Key benefit-risk summary table in the BRAT framework with risk difference.
Source: Coplan PM, Noel RA, Levitan BS, Ferguson J, Mussen F. Development of a framework for enhancing the transparency, reproducibility and communication of the benefit-risk balance of medicines. *Clin Pharmacol Ther*. 2011 Feb;89(2):312-5.

To further enable a direct comparison of benefits and risks, each criterion must be evaluated on its performance. The BRAT framework uses an odds ratio, risk difference, etc. In Figure 8, the results are seen on the right side as a Forest plot. The differences between drug candidate and comparator are clearly seen.

An odds ratio or any other use of a Forest plot is suitable for events, e.g. adverse events and responder rates. For continuous data, e.g. biomarkers, the BRAT framework converts

such data to a dichotomous variable, e.g. percentage achieving a threshold change from the baseline[15]. This is very unfortunate, since any tendencies in data are lost. The BRAT framework does not reveal how very rare events are handled, e.g. the scenario 0 events vs. 1, 2 or more events.

In conclusion, it can be stated that the BRAT framework has a flat structure, which is preferred since it minimises the impact of correlations in a benefit-risk assessment. However, the unreasonable use of dichotomous scoring for continuous data and the inability to handle 0 events vs. 1, 2 or more events is a weakness.

3.6 International Society for Pharmacoeconomics and Outcomes Research (ISPOR)

ISPOR has had a Risk-Benefit Management Working Group (RBM) since 2003. The goal is to evaluate existing methods/tools/models for the benefit-risk assessment of drugs and different stakeholder preferences. The group has recently published a review paper on existing methods[14], but has still not published any developments on new frameworks yet.

3.7 Innovative Medicines Initiative (IMI)

The Innovative Medicines Initiative (IMI) is Europe's largest public-private initiative and is a joint undertaking between the European Union and the European Federation of Pharmaceutical Industries and Associations (EFPIA). PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium) is a collaborative European project that comprises a programme to address limitations of current methods in the field of pharmacoepidemiology and pharmacovigilance.

The overall objective of PROTECT is to strengthen the monitoring of the benefit-risk of medicines in Europe. In order to achieve this overall goal, PROTECT has been designed as a comprehensive and integrated project aiming to develop and validate a set of innovative tools and methods that will:

1. Enhance data collection directly from consumers of medicines in their natural language in several European Union countries, using modern tools of communication;
2. Improve early and proactive signal detection from spontaneous reports, electronic health records and clinical trials;
3. Develop, test and disseminate methodological standards for the design, conduct and analysis of pharmacoepidemiological studies applicable to different safety issues and using different data sources;
4. Develop methods for continuous benefit-risk monitoring of medicines, by integrating data on benefits and risks from clinical trials, observational studies and spontaneous reports, including both the underpinning modelling and the presentation of the results, with a particular emphasis on graphical methods;
5. Test and validate various methods developed in PROTECT using a large variety of different sources in the European Union (e.g. clinical registries) in order to identify and help resolve operational difficulties linked to multi-site investigations[91].

The most interesting activity from a benefit-risk assessment methodology point of view is the work under point 4 concerning the evaluation and development of benefit-risk assessment methods. However, the work is ongoing and no method(s) has yet been published[92].

3.8 Discussion and conclusion of Chapter 3

Current initiatives all indicate that the first step towards transparency is through a structured framework. The major benefit of such a structured framework is that it ensures consistency in decision making. However, a structure in itself is not sufficient, and the effect is limited without clearly defined weighting and scoring of data. Many initiatives are still under

development and it is premature to discuss the pros and cons, but none of the initiatives indicate how they will frame clinical significance and relevance of data, and only one initiative (PhRMA) has a visualisation tool.

The CHMP report from 2008[9] concluded that the MCDA was associated with several concerning features and that a tailored method for the benefit-risk assessment of a drug was therefore much needed. However, the conclusions from the second and third work package in the Benefit-Risk Methodology Project at EMA are that the PrOACT-URL should be adopted and that MCDA could be used for complex decisions[89].

The PrOACT-URL serves the purpose of structuring an assessment. It is extensive in the sense that it systematically takes the assessor through all the important steps in an assessment, where all decisions in the assessment must be justified. The justification aspect is much needed and very important, since it will increase the transparency and consistency in decisions made and will hopefully give creditability to the decision makers.

MCDA in itself is undoubtedly a valuable tool to aid decision making regarding, e.g. political and economic decisions. Several aspects/steps in the MCDA are valuable, but several changes, such as incorporation of uncertainty evaluation related to data of the issues with correlations in the biological system, are needed before it can be fully adapted to the field of medical/clinical benefit-risk assessment.

There is a distinct difference between a decision and assessment. The aim in a decision is to find the optimal decision, while in an assessment the aim is to assess the available information with a full view of correlations, clinical significance, uncertainty, etc. and gather all the information, without compromising the level of information, so that it can aid the decision-making process. In other words, in a decision process it is reasonable to, e.g. add information, while in an assessment all information must be assessed and preserved all the way to the decision makers. A benefit-risk ratio or an overall benefit and risk score are therefore with limited use in an assessment.

The main argument for preserving information all the way to the decision makers is that the value of a single criterion should be able to decide the fate of the assessment. However,

if data are added together, such valuable information is lost due to marginalisation, as discussed previously.

4 Clinical significance

How large does a reduction in disease or symptoms have to be to be clinically significant[102]? Or in other words, what is the smallest change in disease or symptoms that would result in a decision to change treatment[93]? Who defines a meaningful difference? What is reasonable and who defines reason? These are obvious questions in an assessment and seem to be inadequately discussed or defined in current benefit-risk assessment methodologies.

Clinical trials are conducted under strict rules to provide an objective picture of the action of the intervention of interest relative to a comparator. This highly objective data is later processed by statistical departments and is subjectively assessed by project groups/teams and presented to the decision makers.

Boudes emphasises that a major issue is the lack of a discussion of the clinical relevance of data. Randomised clinical trials are designed to answer questions related to the efficacy of drugs/treatments, while safety signals are observed and are therefore the most challenging. Sponsors are often asked to demonstrate clinical relevance of study results, and this is often unsuccessful. The benefit-risk assessments made by sponsors are challenged from time to time by regulatory agencies. Boudes concludes that a better understanding of the conflicting opinions between advisory experts and sponsors will improve the conduct and review of clinical programmes aiming at new indications[25].

After the recent well-publicised regulatory actions related to Vioxx and Avandia, we are faced with a more risk averse environment, and requests for a high standard in the safety profile of a new drug are seen[94].

This has led to increasing costs in drug development, with fewer and fewer drugs getting approved, e.g. by the FDA[95].

Watkins[94] puts the conflicting opinions between advisory experts and sponsors in perspective in a recent paper concerning a new drug for the treatment of a chronic disease. The FDA agreed with the sponsor's assessment of efficacy, but it was noted that two subjects receiving the active treatment experienced abnormal liver biomarkers. The FDA

decided that in order to approve this new drug, the sponsor had to conduct a clinical trial involving 20,000 subjects being treated for 1 year, 10,000 subjects receiving the new treatment and 10,000 receiving a comparator. The economic costs of such a clinical trial, the loss of patent life, and the prospect of losing first in-class position are now facing the sponsor.

Clinical significance and relevance, the backbone of any given benefit-risk assessment, seems to be inadequately defined and appears to be influenced by the individual person, assessor or advisors' experience and expertise. Many current benefit-risk assessments are merely a summary of the statistical results, usually the statistical significant results, and a discussion of the clinical relevance of these findings is often difficult to find. However, guidelines exist that describe the definition of clinical significance for specific disease areas[96] and these guidelines are used when necessary[97]. But the majority of these guidelines are implicit, while explicit in a few cases.

The aim of this thesis is not to evaluate the shortcomings of statistical significance testing in a comprehensive way. However, in order to frame the aspect of clinical significance, a brief discussion of the main advantages, but also the main limitations and shortcomings of statistical significance, is needed.

4.1 Advantages, limitations and shortcomings of statistical inference testing

Statistical and probability science in its modern form can be traced back to 1660, where John Graunt (1620-1674), one of the first demographers, produced the first life table, in which he gave probabilities of survival to each age. Statistics, as we know it today, began to take form around 1750 onwards in the work of Laplace, Gauss, Bayes, Bernoulli, etc., where least squares estimation, probability-based inferences, inverse probability, etc. were developed. Modern statistics found its known form in the 20th century through the contributions of Fischer, Neyman and Pearson[98].

Statistical analysis is used in almost all scientific disciplines to provide credibility to a hypotheses or theory and to condense the data and present the results for interpretation. Most importantly, the development of statistics reflects the need for generalising from a sample of observations on population quantities using a probabilistic argument[98]. This is especially useful when the amount of data are overwhelming.

However, Boudes, Greenstein and McCloskey[25;99;100], just to mention a few, have emphasised that while statistically significant results are essential, there is a lack of discussion of the clinical relevance of data. There is too much focus and reliance on using statistical significance to conclude whether an intervention is clinically significant or not.

A statistically significant result is not always clinically significant; vice versa, a clinically significant result is not always statistically significant. This is explained by the dependence of a statistical significance test on the sample size. In a large sample size, a small difference would be statistically significant, even though the difference is clinically irrelevant. On the other hand, a large clinically relevant difference can be discarded as statistically insignificant if the sample size is too small.

The level of statistical significance, α level of 0.05, often used in clinical trials, is set by Fisher[100]. Hypothesis testing determines whether or not a difference is by chance or is a systematic behaviour. There is 1/20 chance of a type I error; a null hypothesis is incorrectly rejected and is a type II error; a null hypothesis is incorrectly accepted. Statistical significance testing does not reflect the magnitude of the difference, since the main aim is to establish, with a certain level of confidence, whether or not a given difference in mean values is found by chance or is actually a systematic behaviour.

A difference between two interventions is more easily captured in homogeneous populations because of a higher degree of correlation. This means that not only the study population, but also the homogeneity of two populations, will find small statistically significant differences, which are clinically irrelevant. Statistical significance testing provides a yes/no answer based on mean values for a parameter between two populations. As we will see later, mean values provide little information on how treatment effects differ between subjects, and this does not

mean that values give any indication of the proportion of subjects experiencing an effect. The same mean treatment effect can be produced by a small change in all subjects or relatively large changes by a subgroup of subjects and a little change for the remainder[101].

Most of the above criticism relates to Fisherian statistics. An alternative to the Fisherian tradition is the Neyman-Pearson approach first presented in 1928, where they argue that a study should be designed to capture a defined effect size. The magnitude of the effect size depends on how clinical significance is defined by the clinician. In the last decades, this approach has gained recognition, but this method has also been criticised for not providing information regarding the clinical significance of research outcomes[102].

There is a need to capture clinical significance and relevance of data as part of an overarching benefit-risk assessment that can support the statistical analysis. This leads to a discussion of the clinical significance and relevance of data as well as how this can be framed and defined.

4.2 How can clinical significance be framed?

Common sense is dependent on time, person and place, as argued by the famous German philosopher Immanuel Kant (1724-1804)[103]. Søren Kierkegaard (1813-1855), who is seen as the father of existentialism, argues that subjectivity is the truth and that the truth is subjective[104]. It may therefore be meaningless to define universally acceptable rules for clinically relevant differences for each disease area and indications. Since there is no universal or objective definition of clinical significance, several different definitions and approaches to clinical significance exist[100], and in the following the most influential ones will be discussed.

In psychotherapy in the mid-1980s, Jacobson and co-workers discussed and defined the concept of clinical significance, which was a new statistical approach to the treatment-related results in psychotherapy research[105;106]. This approach was founded in a frustration over the

fact that not all statistically significant results were clinically significant. Jacobson and co-workers defined a clinically significant change as the extent to which therapy moves someone outside the range of the dysfunctional population or within the range of the functional population, as seen in Figure 10.

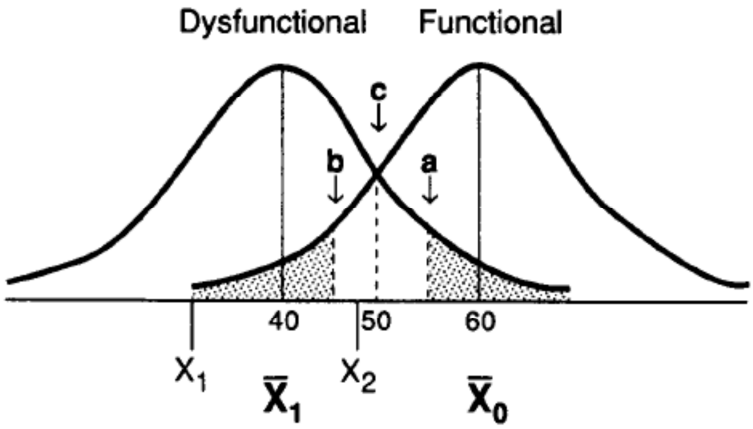


Figure 10: The distributions of the dysfunctional and functional groups are shown. X_1 and X_2 are respectively pre- and post-treatment scores. a, b and c are suggestions for cut-off points for a clinically significant change.
Source: Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol*. 1991 Feb;59(1):12-9.

Jacobson's method is based on the definition of a cut-off point, which defines the boundary between an unhealthy and a healthy patient population. Assuming normal distributions for the healthy and unhealthy populations, the cut-off point may be defined as the point that is halfway between the two means. When functional and dysfunctional distributions are non-overlapping, the interpretation of results is straightforward, since anyone who has crossed the cut-off point must have changed dramatically as a response to the treatment. When distributions overlap, there is a need to determine whether the results are likely or not. To test the result, a reliable change (RC) index is created,

$$RC = \frac{x_2 - x_1}{S_{diff}}$$

where x_1 and x_2 are the pre- and post-treatment values and S_{diff} is the standard error of the difference between the pre- and post-treatment test scores, and it describes the spread of the distribution of change scores that would be expected if no actual change has occurred. The RC index describes how likely a given result is. An RC index larger than 1.96 indicates that there is an actual difference between pre- and post-treatment change ($p<0.05$).

Various other similar methods have since been developed and published and all seem to be comparable[107]. However, there is some criticism of these methods and the main argument is that they are just new statistical tools and that there is a lack of qualitative judgement on data[108].

I agree with Boudes that statistical significance is important, but the clinical relevance of data must be discussed[25]. Jacobson and Truax were not the only researchers frustrated over the limitation of statistical inference. Over the last couple of decades, similar discussions and debates in other fields of biomedical science and health care, e.g. odontology, paediatrics, public health, nursing, etc. have emerged. There has been very interesting debates and discussions concerning the same frustrations that Jacobson and co-workers dealt with[100]. In 1988, Hollon and Flick asked three very essential questions:

1. Should normalisation indices be defined relatively or normatively?
2. How is a meaningful change best defined?
3. Who defines what is meaningful?

They argue that the issue regarding researchers reporting outcomes solely in terms of group means is worrying, since variability within, e.g. the treated group is ignored. Variability has been argued to be at least as important as the mean value. Hollon and Flick suggest that clinical significance or a clinically relevant change depends on the stakeholder. Therefore, they suggest that clinical significance should be defined as the smallest of reliable changes of interest to a given stakeholder, but not necessarily to all stakeholders[101].

In continuation of Hollon's work, Kingman[109], writing in 1992 in the field of oral diseases, argues that a clinically relevant efficacy estimate often becomes defined by default, to be whatever can be demonstrated statistically. If statistical significance is not obtained, this is linked to an inadequately designed study. On the other hand, if a statistically significant result that is clinically irrelevant is obtained, it is also equivalent to a poorly designed study.

Kingman argues that to accomplish a well-designed study, one needs to define a statistically significant rule (SSR), a clinically significant rule (CSR) and a tough clinical rule (TCR). SSR requires that the difference is statistically significant; CSR requires that both SSR is met and that the observed value for the ratio,

$$R = \frac{\mu_T}{\mu_A},$$

is less than 0.90. μ_T and μ_A are mean values treatment T and A for a treatment parameter, e.g. incidence of caries. Values for R between 0.9 and 1.1 (or 90 to 110% expressed as a percentage) by definition represent equivalence between treatment T and A (T=A), values for R less than 0.9 are defined as superiority of T to A (T>A), and values for R greater than 1.10 are defined as inferiority of T to A (T<A). The TCR rule requires not only that SSR and CSR are met, but requires that the two-sided 90% CI for the ratio R is within the superiority range. In Figure 11, a schematic diagram shows the regions of superiority, equivalence, inferiority and least-as-good.

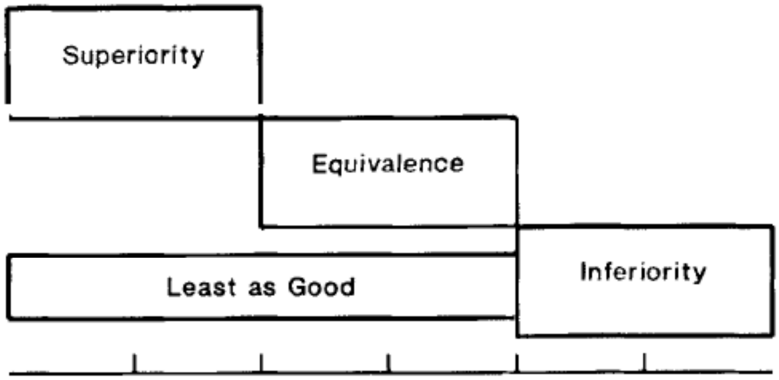


Figure 11: Regions for superiority, equivalence, inferiority and least-as-good regions are defined. Source: Kingman A. Statistical vs clinical significance in product testing: can they be designed to satisfy equivalence? *J Public Health Dent*. 1992;52(6):353-60.

In Figure 12, the SSR, CSR and TCR rules are depicted, and it is seen that the rules are hierarchical, meaning that if TCR is met, CSR and SSR are also met.

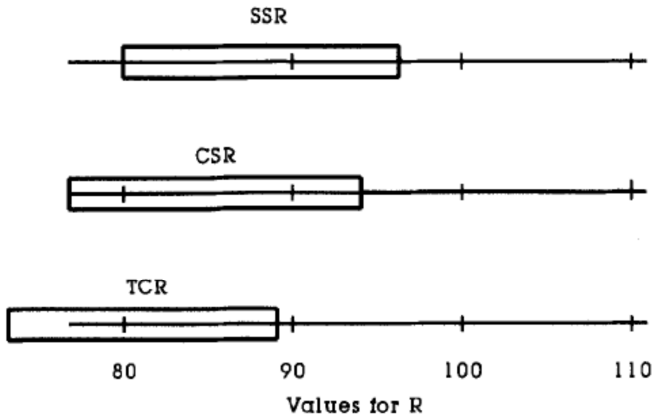


Figure 12: The SSR, CSR and TCR rules. Source: Kingman A. Statistical vs clinical significance in product testing: can they be designed to satisfy equivalence? *J Public Health Dent*. 1992;52(6):353-60.

Kingman suggests that the magnitude of a clinically relevant/significant change is defined in consensus and that a necessary requirement is that statistical significance must be met.

A year later, LeFort[102], a doctoral candidate in the School of Nursing at McGill University, also argues that a summary of statistics (means, SD, etc.) alone does not provide sufficient information about, e.g. a trial, for a clinician to decide whether a new treatment is better, the same or worse. LeFort also expresses many frustrations over statistical significance and the

lack of attention to clinical significance. LeFort suggests that clinical significance must depend on the degree of change, the impact on the individual subject's life, how long the effects last, compliance and cost-effectiveness, etc.

In that same year (1993), Lindgren and co-workers[93], who work within the field of paediatrics, state that the smallest change in a given clinical parameter of interest between two groups that would result in a decision to change treatment is a clinically relevant change. They further state that in practice, clinical significance should be defined prior to the study and that the study should be designed so that a statistically significant change is also considered clinically significant/relevant. Considerations around the sample size are pivotal for a study to obtain both statistical and clinical significance.

Lindgren and co-workers make another very valuable argument: when clinical significance and statistical significance do not agree, the following two unwanted scenarios emerge: 1) a difference is considered clinically significant, but not statistically significant, because the sample size is too small; 2) a difference is considered statistically significant, but clinically irrelevant. This latter scenario indicates that the sample size is so large that even the smallest difference can be statistically significant. This remarkable point that a sample size should not be too small or too large, but just right, may ensure that both clinical and statistical significance can be determined at the same time. Therefore, attention shifts to the power of the trial. This is defined as the probability that a pre-defined treatment difference will be statistically significant, provided that the difference really exists in the population.

Lindgren and co-workers conclude that one cannot decide on a universal cut-off for clinical significance, but rather that the level of clinical significance must be tailored, to some extent, to the disease, disease area and/or the intervention under investigation.

Several years later in 2000, Hujoel and co-workers[110], who are researchers within the field of periodontal science, express concerns about the lack of consensus as to how clinical significance is defined. They suggest the following definition: "*A statistically significant difference in a clinically important outcome as identified in a definitive or phase III trial.*" Clinically important, relevant, or meaningful outcomes are measures of how a patient

functions, feels, or survives, and are specifically defined for each specific problem. The size of the clinically significant difference between treatments can differ between different stakeholders and must therefore clearly be defined and justified.

In 2002, Killoy[100;111], also a researcher in periodontal science, publishes a paper in which it is argued that the definition of clinical significance depends on the individual clinician's judgement and that before any change can be judged as clinically significant, it must be statistically significant. He suggests nine aspects that should be considered when determining the clinical significance of data related to periodontal disease:

1. Efficacy/statistical significance.
2. Improvements ≥ 2mm in probing depth and clinical attachment level.
3. Percentage of sites changing 1, 2 and 3 mm at specific probing depths.
4. Cut-off points, such as reaming pockets ≤5 mm.
5. Percentage of sites returned to health.
6. Percentage of sites still requiring therapy.
7. Morbidity (adverse events).
8. Time to treat.
9. Cost.

Killoy even discusses that frequency distributions may better reflect clinical significance, but does not pursue the idea.

4.3 Building bridges

How do we bridge the obvious gap between statistical significance and clinical significance? In the previous section, different approaches and ideas from different scientific areas, such as psychology, nursing, periodontal science, etc. were presented.

The definition of clinical significance is clearly dependent on time, place, situation and stakeholder perspective (clinician, patient, researcher, regulatory agencies, pharmaceutical companies, payers, etc.). The clinician may consider small statistically significant changes to be clinically relevant, while the patient may mostly consider a treatment successful if there is a specific reduction of symptoms, increased quality of life, etc., and a regulatory agency may be concerned with both the efficacy and safety of a treatment. Clinical significance can be defined by the use of[100]:

1. Absolute criteria: Mean values with a threshold can easily be adopted and implemented, but mean values do not represent variability, which is at least as important as the mean value.
2. Ratios: A ratio can provide a snapshot to compare two treatments. However, there are several concerning features of a ratio. First of all, the same ratio can be obtained with different data-sets, i.e. 4/2 = 2 and 16/8 = 2. Secondly, information is lost once data are reduced by dividing, adding and multiplying. Thirdly, ratios can be very misleading, because a two-times better result may represent only a small change in the effect size when comparing two treatments, e.g. weight decrease of 100 grams for treatment A and 200 grams for treatment B.
3. Frequency distributions: They reflect specific results about a parameter/variable at specific values, e.g. how many patients lost 1 kg, 2 kg, 3 kg, etc.
4. Cut-off points: A specific threshold can be used to define a clinically significant change for a given variable/parameter, e.g. a decrease in HbA1c to less than 7% is considered clinically relevant and significant. The downside of using cut-off points is that data are dichotomised, which is very unfortunate for continuous data, since a great amount of information is lost, e.g. variability, tendencies, etc.
5. Percentage of improvements (proportions): In the comparison of two treatments, the percentage of improvements and the associated distribution show the variability of data, which otherwise will be lost when a cut-off point is defined. A threshold can be

defined describing the clinically significant and relevant percentage of improvements that are required to claim a difference between two treatments. A downside to the definition of cut-off percentage of improvements is that even the smallest treatment effect will count as an improvement. This can be accounted for by defining a percentage of improvement more than a specifically defined level, e.g. in an obesity trial, it could be the percentage of improvement of more than 2 kg of weight loss.

6. Percentage of patients affected (proportions): This is more or less the inverse of the percentage of improvements.
7. Disease progression inhibition: This is in line with “percentage of improvements and patients affected”, but the focus is on inhibition of disease progression. Stakeholders need to clearly define the disease and a threshold for disease progression, which might be arbitrarily defined.

4.4 Discussion and conclusion of Chapter 4

Classical statistical inference testing has its advantages and restrictions, which has led to the development of other methods, e.g. clinical significance testing by Jacobson and co-workers[105;106]. However, these methods are restricted in their use because of the assumptions made, e.g. that data are normally distributed, etc. There are several other concerning issues related to Jacobson’s proposal: the dysfunctional and functional distributions overlap; how do we distinguish dysfunctional cases with extremely “normal” scores from non-dysfunctional cases? How do we decide what constitutes a functional or dysfunctional population[101]? The clinical significance method developed by Jacobson is conservative in its use, but the idea of defining clinical significance was a breakthrough for methods, models and ideas to follow.

Descriptive methods, such as percentage of patients improving, frequency distributions, etc., provide other means in the assessment of clinical significance of data. To ensure a transparent and relevant assessment, clinical significance must be defined and justified from

assessment to assessment. As Greenstein[100] emphasises: *"Clinical trials are conducted to answer clinical questions, and clinical parameters are used to monitor outcomes; therefore, the results should refer to the importance of the clinical data…… Ultimately, it is incumbent on investigators to address the clinical importance of their data to help clinicians choose effective therapies, because reliance on hypothesis testing to define 'clinically significant findings' is inadequate."*

In Chapter 5, the benefit-risk assessment method developed during the PhD project will be presented. How clinical significance can be perceived and how this can be captured in clinical data will also be discussed.

5 A benefit-risk assessment approach

The aim and purpose is - in a simple and structured fashion - to compare benefits and risks on the same scale, and thereby support decisions under drug development and in a marketing authorisation application. It is therefore important that the assessment is as transparent as possible and that focus is on the clinical significance and relevance of data.

The development of a novel methodology and visualisation tool for the benefit-risk assessment of medicines and treatments will be elaborated. Previously, several different methods and tools were discussed. Relevant aspects of these methods are taken into consideration and built upon to develop a tailored method for the benefit-risk assessment of medicines and interventions.

The goal is not to create a decision model, which gives a go/no-go answer, but a tool for the decision-making process. Subjective input into the assessment is needed, but stringent rules are applied to retain transparency in the assessment. The main focus in this thesis will be on the general framework, scoring of data and visualisation and communication of the result. The technical part, where data-driven methods are used for scoring, will be described, but the focus will be on how the results obtained by these technical methods can be used in a clinical setting and in a benefit-risk assessment.

The proposed methodology was developed as an iterative process, starting with a simple approach that was adjusted and built upon continuously. A preliminary method was firstly developed. A suitable project was identified and a project team was invited to a workshop to test the method. The project team was asked to identify which type of data they would like to include in the assessment. Often they chose 2-3 phase II and III clinical trials.

Every criterion was identified and listed. This included not only primary and secondary endpoints, but also, e.g. biochemical markers. The new drug was scored against the comparator using either qualitative or quantitative tools.

A workshop was then organised where the entire project team was involved, that is to say the project manager, the international medical director, the regulatory affairs associate, the

statisticians, the medical writer, the safety surveillance advisor, the clinical pharmacologist and the medical affairs advisor. The team was then presented to the benefit-risk assessment method and was later asked to define the decision context, criteria and assign weights. We then provided the objective scores and the team could revise the subjective scores. An evaluation of uncertainty was performed and the final weighted scores were presented using different visualisation tools. Finally, the team made the final conclusion of the assessment. The team then evaluated the process and this valuable input was then used to adjust the method. This process was repeated, as mentioned previously, with four different drugs and a different project team. New issues emerged each time that had to be dealt with, hence the method was developed in an iterative fashion.

As illustrated in Figure 13, the proposed benefit-risk assessment method is structured as an eight-step successive process. This structure serves the purpose of transparency in the process and it contributes to reducing the effects of unintended bias and feedback. In the following, the rationale for each step will be described. Basically, the eight steps can be divided into 3 main groups:

1. Introductory steps in an assessment
2. Evaluation of data
3. Visualisation and communication of results

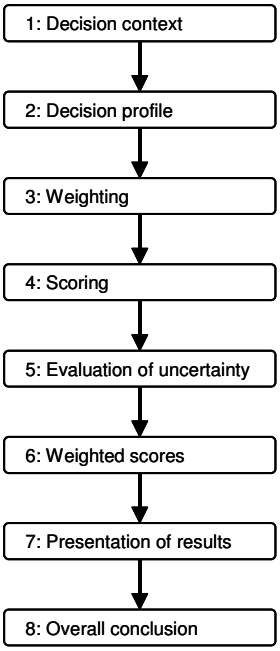


Figure 13: The eight successive steps in the benefit-risk assessment method.

5.1 Introductory steps

The first steps in the assessment are pivotal and lay the boundaries for the remainder of the assessment. Without clear boundaries for the assessment, the decision at the end will lack transparency and credibility. The purpose is therefore to ensure that every stakeholder will be able to understand the motivation, aim and goal of the assessment.

5.1.1 Decision context - Step 1

The first aspect that needs attention is the decision context. The decision context is the most important step in a benefit-risk assessment. As mentioned in the evaluation of MCDA in Chapter 2, there is the risk of performing a perfect analysis for the wrong problem, if the wrong question is answered. However, the decision context is dynamic and can be changed throughout the assessment as new issues emerge, which may indicate a change of the aim. A clear aim helps to keep the assessment on track[19]. The aim and goal of the assessment must therefore be clearly defined.

There seems to be consensus in the literature that a benefit-risk assessment should always include comparators[16]. The comparator must therefore be clearly defined in the decision context. This could, if none are available, be *"no treatment"* or a placebo. The benefit-risk assessment can include two or more drugs. The drug(s), comparator(s) (active comparator and/or placebo) and dose(s) to be included in the benefit-risk assessment are clearly defined. It is also clearly defined who is to be involved in the assessment. A specific benefit-risk assessment is made for each indication[5;9;16;18;80].

The decision context lays the boundaries within which the decision is taken. If the decision context is concerned with the consideration of a decrease of HbA1c and how this correlates to hypoglycaemia, then the number of criteria in the disease/indication model is initially at least limited to HbA1c and hypoglycaemia. Every aspect of the decision context must be described[5;18].

Another important aspect of the decision context is the definition of clinical significance. By defining this beforehand, unintended bias in the decision-making process is prevented and the assessment will gain credibility. As discussed in Chapter 4, the definition of clinical significance/relevance varies from context to context. Therefore, there is no universal definition of clinical significance/relevance. However, as Lindgren discusses[93], one needs to define the smallest change in a given clinical parameter of interest between two groups that would result in a decision to change treatment. In this thesis, it is more natural to define and discuss clinical significance in step 4 (Scoring). Once the decision context is clearly defined, the next natural step is to identify criteria that frame the decision context. This is Step 2 in the process.

5.1.2 Decision profile - Step 2

The benefit and risk criteria that characterise the decision context are identified. The most relevant criteria are then selected and justified. The specific selection of criteria to be

included in the assessment must be justified not only for documentation and communication purposes, but also to ensure transparency in the assessment.

A benefit criterion in one assessment can sometimes be a risk criterion in another assessment. A strict definition of criteria into benefits and risks is not necessary, but it seems to be a habit from the past. Therefore, most importantly, one needs to define and put into context which criteria are relevant and why, and what are the wanted and the unwanted treatment effects, etc.

Justification not only supports transparency, but also enables one to trace back decisions made in the past. This is important since the development of a drug takes about 10 years. During this time, people move around and it can therefore be difficult to keep track of the reasoning behind all decisions made. If a benefit-risk assessment report is conceptualised at the beginning of a project, and updated continuously as data emerges and decisions are made, a natural history of decisions will be preserved.

A criterion is a specific parameter that describes an action of a drug. The selected criteria are common for all drugs and comparators in the assessment, because all drugs in the assessment must be compared on equal and fair terms. So, a problem arises when some criteria are formulated such that they are only valid for one of the drugs. An example could be *injection site reaction at administration*. This criterion is obviously not valid for drugs administered orally. Therefore, such criteria must be reformulated so that they are relevant for all drugs to be assessed. In this example, the criterion can be formulated as *side effects from administration.* If possible, each criterion should be based on a measured parameter in a clinical dataset.

As expressed by CHMP[9], the loss of information should be minimised. Any grouping of criteria should be avoided, since it is meaningless to assess overall SAE, because any SAE that stands out will simply vanish in such groupings.

Not all criteria can be measured on performance because of different factors, e.g. time limit[9;16]. Such criteria can also be included in the assessment and their performance is obviously subjective and must be justified. An example is a preclinical rare cancer risk

finding, and the obvious concern is whether it also applies to humans. Since it will be very time consuming and expensive to invest in clinical trials before marketing, to clarify this question, it can be included in the assessment and weighted and scored subjectively. This leads us to the next step in the assessment, "*weighting".*

5.1.3 Weighting - Step 3

All criteria are not equally important in a benefit-risk assessment. Therefore, it is reasonable to define a relative importance or weight for each criterion, indicating how important the criterion is on a value scale[17;19]. Furthermore, benefits and risks, which are different entities, can only be compared on the same scale if they are weighted on the same value scale. The weights are based on the importance of differences and clinical judgement, or in other words, the weight of a criterion is based on the relative importance of a difference between two treatments, relative to the importance of a difference for other criteria within the given decision context. Weights also reflect the fact that a difference between two treatments will have clinical implications depending on the importance of the criterion. Thus, weighting and clinical significance are closely related.

Several methods exist for assigning weights, e.g. swing weighting, hierarchical methods, non-hierarchical methods and rank-based methods[17;80]. It is not the purpose of this thesis to give an elaborate account of these methods. However, I will shortly describe the most favoured method of "swing *weighting"*, which defines a scale from the worst preference to the best preference for each criterion. The scale is defined as being from 0 to 100, where 100 is the most favoured value.

In the use of swing weighting, three scenarios emerged: In the first scenario, it was very difficult to obtain consensus in the research group. A scale from 0 to 100 led to a daunting task, especially when many criteria were chosen. The weighting was influenced by large inter-subjective differences and many endless discussions between individuals in the group. In the second scenario, I experienced that people tended to divide the scale from 0 to 100

into three of four equal intervals, e.g. when asked to assign weights based on preference, the following weights were given: 0-25, 25-50, 50-75 and 75-100 or 0-33, 33-66 and 66-100.

In the third and final scenario, it was observed that when people were knowledgeable about the result from clinical trials, which they were assessing, the weights were influenced by their knowledge. Based on these findings, it was concluded that a broad scale would not necessarily result in a more transparent weighting and that weighting must be conducted before data are revealed to avoid any unintended bias.

A scale from 1 to 3 was defined and tested with satisfactory results. Each benefit and risk criterion is assigned a weight/importance of low (1), medium (2) or high (3). The weights are comparable across all criteria and drugs in the assessment, meaning that two criteria with the same weight are equally important. Weighting must be independent of the data sets, meaning that they are assigned before datasets are revealed. This ensures that the weights are not biased. As described above (the third scenario), it can easily be imagined how weights can be influenced by knowledge of results. Criteria that perform undesirably may be weighted low, and unexpected positive results may be weighted high, e.g. unexpected decrease in blood pressure in a weight-loss trial.

The final weights can be based on consensus between the benefit-risk assessment group and the regulatory agency. In preliminary meetings between agencies and the sponsor, the weights can be decided in consensus. This collaboration would ensure that the final benefit-risk assessment by the sponsor in connection with a marketing application will have credibility. As with the selection of criteria, all weights must be justified for documentation and communication purposes, but more importantly to ensure transparency in the assessment.

The simple weighting scale improves transparency and leads to discussions of importance instead of focusing on a number, e.g. should the weight be 60 or 65% for a given criterion? Different stakeholders are likely to have different preferences and they may also disagree about the choice of criteria. It is therefore important that this information is fully maintained and clearly presented in the final decision process.

It is also important to notice that the weighting process in this thesis has shifted focus from preference, which is the basis for weighting in MCDA, to the importance of a difference between treatment A and B for a given criterion within a given decision context. While weighting is concerned with importance, the focus is now shifted towards performance, which is elaborated in the next section.

5.2 Evaluation of data

The evaluation of data is divided into two main steps: scoring and evaluation of uncertainty and evidence. The evaluation of data puts the performance of a drug and comparator into perspective and is also closely related to the definition of clinical significance. Therefore, we are forced to broaden our understanding of data and the clinical significance and relevance of differences between two options. In addition to the statistical analysis, scoring of data and discussion about clinical significance and relevance of data will bring new light to the assessment.

Different types of data require different measures for scoring, but most importantly clinical significance must be defined. In Chapter 4, it was discussed how clinical significance could be framed. In the following, the learning from Chapter 4 will be modified and applied. Another aspect in the evaluation of data are to capture tendencies in data. This is important in order to protect patients from potentially harmful adverse effects of a drug. A proactive approach is needed, where willingness to accept a false positive signal overrules the risk of accepting a false negative signal. This is elaborated under section 5.2.1.2 Confidence interval scoring.

The evaluation of uncertainty and evidence in data different measures, such as non-parametrical and parametrical re-sampling, will be discussed.

5.2.1 Scoring - Step 4

Scoring is the process of assessing the performance of the drug in the assessment against the comparator for each of the selected criteria[17]. The performance of each criterion is

translated into a score, whereby benefits and risks are brought into a common scale and can be added and subtracted, and can be discussed with regard to their significance in a qualitative manner.

Scales for the scoring process must be constructed to enable the scoring of criteria and many subjective scoring methods exist, as described in the literature[5;12;16;80]. Two scenarios were evaluated; should the scales be constructed before or after the clinical trial data were received? In the first scenario, scales were constructed beforehand to avoid unintended bias. However, this led to undesirable large scales and the scores were more likely to lie close to one another in the middle of the range, which made the discrimination between the drug and comparator difficult.

In the second scenario, the scales were constructed after clinical trial data were received. The scales were then very narrow and the scores were more likely to be far apart. In this situation, a small difference between drug and comparator induced large and unreasonable differences, such that it affected scores between the drug and comparator involved in the assessment.

In the creation of scales, in whatever method was chosen, the variability in data were not handled very well. In an attempt to create scales, SD (standard deviation) was taken into consideration. What does SD essentially tell us, apart from that there is variability in data? Many subjects in a trial could have a low variability, but few subjects with a large variability could be the main driver in the creation of an overall large SD for a given criterion. It does not provide a much needed overview.

Another important goal was that the scales must correspond to the subjects in the trial. Every trial is different from one another, e.g. different countries, ratio between sexes, mean age, batches used, healthy vs. patients, disease stage, etc. Therefore, there is a need for custom-made scales for each clinical trial. However, one cannot simply create scales based on baseline values, because any change within a category is not captured, e.g. if a category is defined as 5-7% HBA1c, a subject that decreases from 6.9% at the beginning of the trial to 5.1% at end of the trial will still be in the same category, and thus no movements are

detected. In conclusion, there is a need to look upon raw clinical trial data for every single subject to achieve and create the basis for scoring.

Furthermore, there is also a need to capture the clinical significance and relevance of data in the scoring. As it was discussed in chapter 4, there are several ways to do this. From a clinical point of view, it is proposed that clinical significance should be based on the proportion of patients experiencing a wanted treatment effect, or in other words by modifying Lindgren's statement [103]: what is the smallest difference in the proportion of patients for a given clinical parameter of interest between two groups that would result in a decision to change treatment?

Clinical significance must be defined and justified from assessment to assessment based on the decision context, which therefore needs to be clear and crisp, leaving no room for speculations or questions. When dealing with large sample sizes, the aim is to determine the proportion of patients experiencing a treatment effect. From a clinical point of view, the higher proportion, the better. A cut-off point can be determined for each disease or disease area based on the decision context, medical knowledge, etc.

It is obvious that this depends on a large number of patients in a given trial; however, when dealing with small sample sizes, the aim is to capture any tendency in data that can later be quantified in proportions in larger sample sizes, as described above.

Different types of data require different methods for scoring. Two main methods have been developed that can handle most clinical data. The methods are objective, relying only on a few subjective definitions, and it is here that the data itself determines the results. The smaller sample size, the more uncertainty there is related to the estimate that is made. Depending on sample size, two strategies were developed, as will be elaborated in the following.

A numerical value is assigned to each criterion, based on available datasets from clinical trials or other information, e.g. preclinical data. For each criterion, the drug is scored relative to the comparator on a simple and transparent scale: -1 (inferior), 0 (non-inferior) and +1 (superior), as seen in Table **6**.

Table 6: The drug is scored against the comparator for each criterion. The drug score can be +1 (superior), 0 (non-inferior) or -1 (inferior).

Criterion	Score
Drug is ***superior*** to comparator	+1
Drug is ***non-inferior*** to comparator	0
Drug is ***inferior*** to comparator	-1

In the following, each strategy is described in detail for the different types of data that is encountered in a clinical dataset. For continuous data and frequent events, a scoring method - difference distribution scoring - based on proportions is used. For events like rare adverse events or responder rates, confidence interval scoring is used to capture tendencies.

5.2.1.1 Difference distribution scoring

Clinical significance must be reflected in the scores. Instead of developing a new statistical inference test, a pragmatic approach using descriptive statistics is favoured. From a clinical point of view, statistical significance is inferior to clinical significance. The previous section proposed that clinical significance should be defined as a given proportion of the subjects in a trial experiencing the desired effects of a treatment.

A clinically significant difference may be described, e.g. by using a 2:1 principle. This means that two out of three patients have to show a difference for a given criterion relative to a comparator, in order to receive a score different from 0 (non-inferior). This is thus defined as a clinically significant and relevant difference.

Different ratios may be used for different indications. The fundamentals behind this principle are to compare fractions of data. The principle is independent of the number of patients in a trial, although the associated uncertainty increases with smaller sample sizes. A simple scale (-1, 0, +1) is chosen and the main purpose of this simple scale is to clearly

reflect the uncertainty that is related to assignment of scores. I do not think that any broader scale is justifiable. The 2:1 principle is demonstrated by the following example:

> **Example:** Imagine a cross-over trial investigating the effect of a new antihypertensive drug (Drug A) relative to an already marketed antihypertensive drug (Drug B) on the effect parameter of systolic blood pressure. The subjects are first given Drug A and then Drug B and the end-of-trial values for systolic blood pressure for Drug A and B are subtracted for each patient and plotted. The resulting distribution is the difference distribution for the two end-of-trial distributions. If most values are bigger than 0 (A - B ≥ 0 mmHg), we thus conclude that Drug B is more effective than Drug A, since a decrease in systolic pressure is a wanted effect. The difference distribution shows in percentage the proportion of subjects experiencing either a better or worse effect with a given treatment.

Randomisation ensures that the groups intended for comparison are perceived as similar. This can be extrapolated to the statement that two groups can be perceived as the same group, receiving two different treatments, just like a cross-over trial. However, since we are dealing with different groups of subjects, the difference distribution cannot be directly created and is instead either created empirically or based on fitted distributions[112]. The aim of this thesis is not to give a mathematical evaluation of difference distributions and how they can be derived mathematically, but how these difference distributions can be used in a clinical setting.

For continuous variables, such as biomarkers (e.g. blood glucose), vital signs (e.g. blood pressure and weight) and frequent events that can occur more than once per subject (e.g. minor hypoglycaemic events), the scoring method is based on difference distributions.

The 2:1 principle is not necessarily universally applicable to all disease areas, but the main argument is that whatever ratio is chosen, the scoring is data-driven and all changes from the baseline are captured. For each drug and comparator, the selected criteria are assigned a given score. The performance of each drug against the relevant criteria is assessed, based

on relevant data seen as changes from the baseline or as compared to the placebo/comparator. The choice of a clinical significant difference is justified and is optimally done under the decision context prior to data analysis.

Continuous data

Continuous data, e.g. biomarkers, represent criteria that are measured for each patient throughout a trial. In most cases, these parameters are measured on a continuous basis, and there is normally a start-of-trial value and end-of-trial value. From a clinical point of view, it is interesting to describe the proportion of patients experiencing a treatment-related effect, the magnitude of the effect and the distribution of patients as a function of the magnitude of a treatment-related effect. In other words, who is experiencing an effect, and how big is the magnitude of the effect?

In Figure 14a, the criterion “weight change” is shown. The blue dots represent end-of-trial data from the new drug and red dots represent the comparator. In these cases, data are fitted to normal distribution for drug and comparator and the two distributions are used to create their difference distribution. Figure 14b shows that 73% of the subjects in the drug group experienced a larger decrease in weight, compared to the subjects in the comparator group, or in other words, 23% more patients in the drug group experienced a weight loss, which they wouldn’t have if they were given the comparator. Based on one’s definition of clinical significance, it is decided whether or not the difference is clinically relevant.

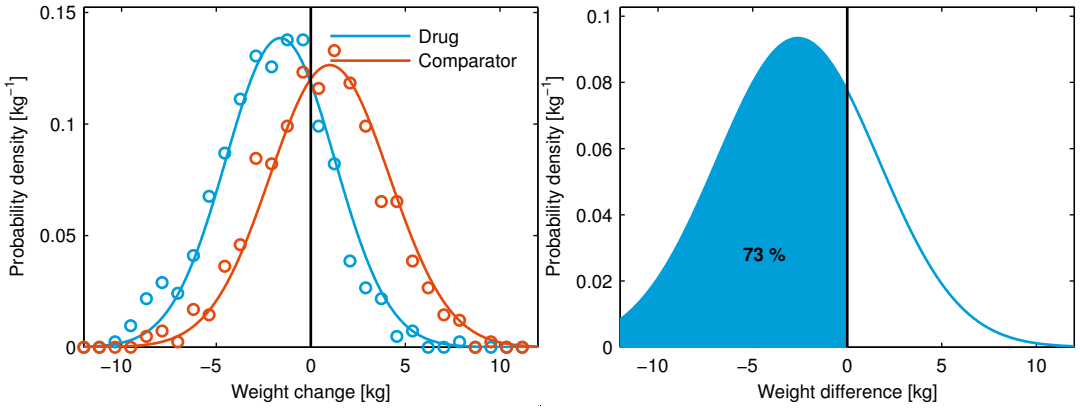


Figure 14: In a) end-of-trial data for the drug and comparator are shown. Data are in this case fitted to normal distribution. The two distributions are used to create their so-called difference distribution, plotted as seen in b). 73 % of the subjects in the drug group experienced a larger weight loss.

Whenever possible, difference distributions should be created empirically. Any assumption about the distribution of a dataset can always be questioned. With empirically created difference distributions, there is still the issue of uncertainty that increases with a declining number of subjects in the analysis[113].

Figure 15a shows progression-free survival of two different chemotherapy treatment regimens[114]. Data are not fitted and typical survival curves are seen. In Figure 15b, their difference distribution is shown. It is seen that 63% of the patients in group II had a longer progression-free survival compared to group I.

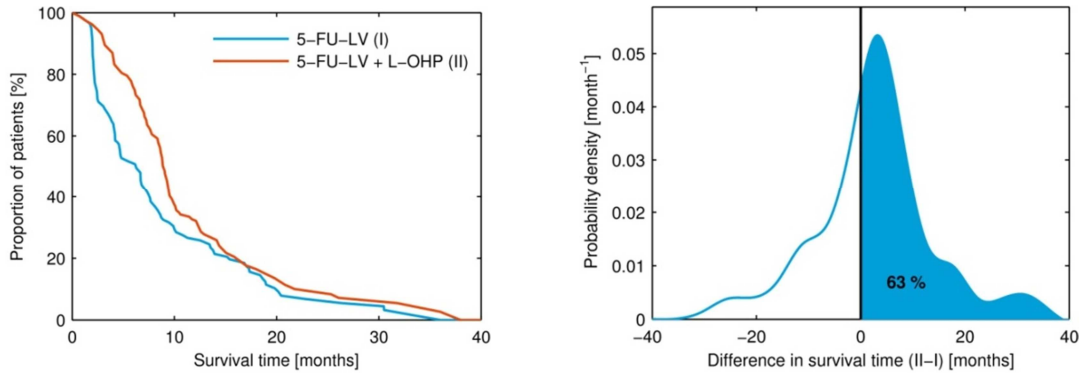


Figure 15: a) End-of-trial data for two different chemotherapy treatment regimens are shown. Data are not fitted in this case. b) 63 % of the subjects in group II experienced a longer progression-free survival time, compared to treatment I.
Source of Figure 15a: Giacchetti S, Perpoint B, Zidani R, Le BN, Faggiuolo R, Focan C, et al. Phase III multicenter randomized trial of oxaliplatin added to chronomodulated fluorouracil-leucovorin as first-line treatment of metastatic colorectal cancer. *J Clin Oncol*. 2000 Jan;18(1):136-47.

In some cases, there is no comparator available or a placebo might be unethical. Under these circumstances, the untreated disease itself may provide baseline values. For the efficacy parameters, usually biomarkers or other biophysical parameters, the baseline values can be used to create the difference distributions once the end-of-trial values for the drug have been obtained.

Frequent events

These events may occur more than once per patient throughout a clinical trial. For frequent events, e.g. minor hypoglycaemic events, the same method is applied, but data are in this case fitted to exponential distributions, which often offer an appropriate description. As with continuous data, from a clinical point of view, it is interesting to capture how many patients experience these frequent events, and how often. A clinical judgement can then be made to distinguish between two treatments, based on a definition of what the smallest acceptable difference is.

Figure 16a shows the number of minor hypoglycaemic events per patient for the drug (blue) and the comparator (red). The two distributions are used to create their difference distribution (Figure 16b), where 71 % of the subjects in the drug group had fewer minor hypoglycaemic events than with the comparator.

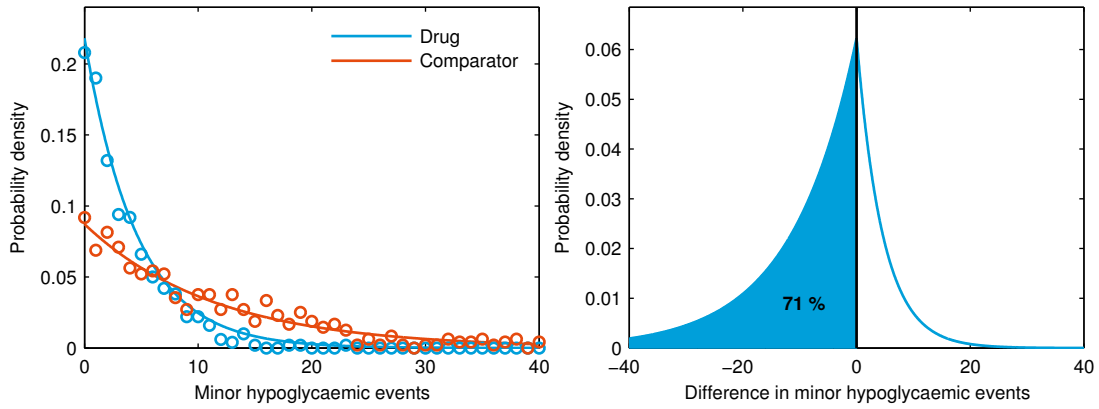


Figure 16: a) End-of-trial data for the drug and comparator are shown. Data are in this case fitted to exponential distributions. b) The difference distribution is shown. 71 % of the subjects in the drug group experienced fewer minor hypoglycaemic events.

5.2.1.2 Confidence interval scoring

As mentioned in the beginning of section **5.2 Evaluation of data**, another aspect in the evaluation of data is to capture tendencies in data. Randomised clinical trials are often constructed so that they answer questions related to the efficacy of a drug. Adverse events are often summarised without any statistical analysis. In order to protect patients from potentially harmful adverse effects of a drug, a proactive approach is needed, where willingness to accept a false positive signal overrules other decisions. One is faced with the two options: (i) the risk of pursuing a false positive signal that will have economical costs without a doubt, or (ii) the risk of ignoring a signal that might have both economical and human costs.

For events (e.g. responder rate, rare adverse events, etc.), the scoring method is based on confidence intervals (CI)[113]. During a trial, not all subjects will experience an event. Therefore, a value is not obtained for each subject, and we can therefore not reuse the method described under difference distribution scoring because probability distribution function (PDF) or cumulative distribution function (CDF) curves cannot be drawn or justified. Instead, it is assumed that such events occur independently, with the same probability for each subject. This is assumed, since it is expected that randomisation will ensure similar groups. The confidence bounds, calculated by the exact method of Clopper and Pearson[115],

are used for scoring. This is a well-known and established method, and I will therefore focus on how this can be used in a clinical setting to capture tendencies in data that can be quantified in larger trials.

A chi-square test and Fisher's exact test were also evaluated and the following was concluded: a chi-square test cannot handle small numbers and the case of 0 events and was therefore discarded. Fisher's exact test is suitable for small numbers and can handle the case of 0 events, but the method calculates a p-value, which was not found suitable in the comparison of two options, since it can be difficult to interpret a p-value. Clopper-Pearson's exact method can both handle small and large sample sizes and the case of 0 events.

Rare events are defined as true or false, e.g. serious adverse events (SAE) leading to withdrawal from the trial, a responder rate, etc. An event is assumed to occur only once for each subject in the trial period, or at least the probability of the event occurring twice is negligible. Furthermore, we assume that such events occur independently with the same probability for each subject. The probability is assumed to be dependent only on the *treatment,* and the number of events in a trial is hence assumed to be binomially distributed. Confidence intervals (CI) are used for scoring, e.g. in the case of events. The number of events and subjects in each group is known. The question is whether the probability, *p*, of one event/subject is different between drug and comparator. By calculating a score for each possible scenario, a scoring table is created, as seen in figure 17. A score can hereby easily be determined for the drug for a given criterion based on clinical data.

In Figure 17, a hypothetical trial, including 500 subjects for both the drug and comparator group (the method is not restricted to the same number of subjects in both groups) is shown. The drug is inferior to the comparator on the criteria headache and anorexia, but non-inferior on the criterion injection site reactions.

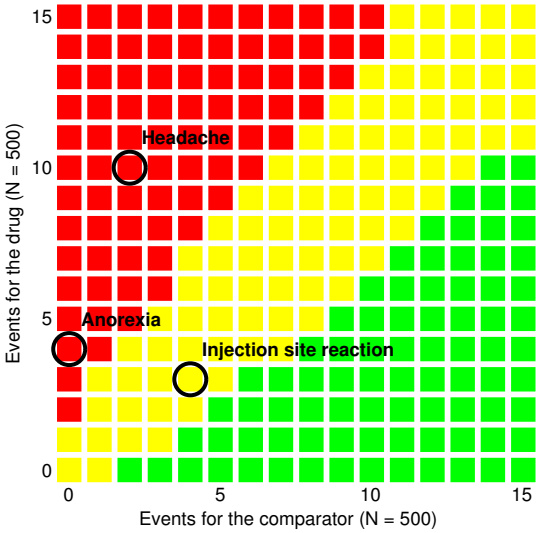


Figure 17: The scoring table for a hypothetical trial is shown. Each little square is the result of a confidence interval (CI) scoring. The red colour means that the drug is inferior to the comparator; yellow means the drug is non-inferior and green means that the drug is superior. There were 10 events of headache in the drug group and two events in the comparator group. The drug is therefore considered inferior to the comparator on this criterion. Furthermore, the drug is considered inferior on the criterion anorexia and non-inferior on injection site reactions.

To be able to capture tendencies in data, a proactive approach is taken, in that the level of confidence is lowered. This will, as mentioned, undeniably increase the rate of Type I errors; therefore, any signal detected must therefore be evaluated and if there is any indication of relation between the action of the drug and the adverse event, it must be investigated and quantified in future studies. As more data are gathered, the level of confidence can be raised to minimise the risk of Type I errors.

Some events occur more than once during a clinical trial, but not in the same manner as frequent events. Here a modified version of the above-mentioned method could be used in these situations. It can be assumed that during a very small time frame, all patients have the same probability of experiencing an event. Furthermore, it can be assumed that the probability is constant over time.

In situations where no comparator or placebo is available, the background incidences for the untreated disease can be used as baseline values, both in the cases where new risks are introduced by the drug and in cases where the considered types of risk already exist in the untreated state, but are promoted by the drug. Once the objective scores are determined, the

next natural step is an evaluation of the evidence and uncertainty surrounding these scores, and this leads to the next step.

5.2.2 Evaluation of uncertainty - Step 5

The objective scores must be reviewed critically. CHMP[9] recommends that the following aspects can be evaluated subjectively, as part of a qualitative evaluation:

1. Discuss the choice of dose, comparator, and endpoints (including surrogates, as appropriate).
2. Describe important methodological flaws or deficiencies. Refer to guidelines or scientific literature if useful and describe how deviations from guidelines or scientific advice, if any, have been justified.
3. Describe the impact of methodological deficiencies on the estimated benefit, e.g. consider any issues of multiplicity, exploratory techniques, post-hoc analyses, etc.
4. Have measurements and scales been validated? What are the unsettled issues? Is there a need for further studies?
5. Describe any negative studies and studies showing no difference.
6. Describe the quality of the supportive scientific literature.
7. Describe any other issues that may have an impact on the estimated benefits.
8. Are the results consistent across different factors, e.g. pivotal trial(s) and supportive studies, all submitted studies and literature, different populations, centres, doses, etc.?

Uncertainty related to the data-driven scores can also be estimated quantitatively on data level, e.g. by non-parametric re-sampling techniques, using bootstrapping[71;116]. This can support the qualitative evaluation, but should not be a stand-alone procedure. Some trials have few subjects involved and others have thousands involved. The fewer subjects

involved, the weaker the power of the trial. This is incorporated in the scoring by using re-sampling techniques. When creating the difference distribution, as seen in figure 18, the power of the trial is depicted as the blue band instead of a single line. If the band is not completely clear of the 2/3 bound, it cannot be concluded that the score is either -1, 0 or +1. The score is then an interval, e.g. 0 to +1, meaning that for a given criterion, the drug is non-inferior to superior relative to the comparator. figure 18 shows that for the criterion “weight loss”, where the number of subjects is 40 in both groups, the drug was not completely clear of the 2/3 bound, hence the drug scored 0 to +1.

In my experience, people tend to focus on numbers rather than a qualitative judgement. This could potentially have undesirable consequences for decision making. The problem is often illustrated by the misuse of “p-values”, where a p-value less than 5% is accepted and the difference is considered real, while a p-value bigger than 5% is rejected and no difference is claimed. How big is the difference between two p-values of 4.9% and 5.1% respectively?

A major drawback with non-parametrical re-sampling is that it cannot handle the case of 0 events in population. Regardless of how many times the re-sampling is performed, an event will never be obtained. This issue can be handled by parametrical re-sampling, where data are fitted to a distribution, and the re-samples are drawn from the distribution. However, the assumptions behind the distribution can always be questioned and the credibility of the re-sampling is hereby challenged.

I believe that the qualitative evaluation of uncertainty and evidence is the cornerstone of an evaluation of data and scores. A qualitative evaluation can be supported by quantitative measures when needed. The evaluation of uncertainty is integrated in the data-driven scoring method in a simple way. In the case of any uncertainty, due to one or more of the above-mentioned aspects, the score may be given as an interval. The interval -1 to +1 should be chosen, when the available information can only be regarded as trends. Any assignment of interval scores must be justified.

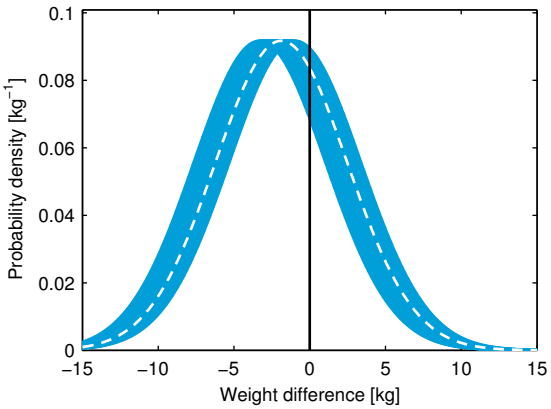


Figure 18: The criterion weight loss from a trial. The dashed line indicates the 2/3 bound and the re-sampled difference distributions are shown as the blue band (2/3 two-sided confidence interval). The re-sampled cases are on both sides of the 2/3 bound and the drug score is 0 → +1.

5.2.3 Weighted scores - Step 6

Previously, the crucial impact of weighting on a benefit-risk assessment was discussed. Weights allow prioritisation of criteria, but say nothing about the performance of the medicine. Scores, on the other hand, capture the performance of the drug, but say nothing about the clinical importance of a given difference in performance. The weights and scores are multiplied to produce weighted scores. Many of the methods and tools described in Chapter 2 combine the importance or preference of a criterion with its performance.

The crucial role of weights is to bring the different scores onto a common weighted score scale, which enables the direct comparison of different criteria. The resulting weighted scores can become -3, -2, -1, 0, 1, 2, 3 or an interval -1→0, 0→1, -1→1. In Table 7, we see examples of criteria, their weights, scores and weighted scores.

Table 7: Example of criteria, their weights, scores and weighted scores for a hypothetical medicine.

Criterion	Weight	Score	Weighted Score
HbA1c	3	0	0
Fasting blood glucose	3	+1	+3
Major hypoglycaemic events	2	+1	+2
Weight loss	2	0→+1	0→+2
Minor hypoglycaemic events	2	0	0
Injection site reaction	2	0	0
Convenience	2	+1	+2
Anorexia	1	-1	-1
Headache	1	-1	-1

The weighted scores cannot only be directly compared, but since criteria have been brought onto the same scale, they can be added, multiplied and discussed in a qualitative manner. However, it is important to bear in mind that total weighted scores, average weighted scores etc. do not contain the same amount of information, as seen in Table 7 it can, though, be valuable to calculate average weighted scores, when the benefit-risk assessment, e.g. contain multiple similar trials and there is a need to reduce the amount of information, enabling a more approachable visualisation of results. This is further elaborated in section 5.3. Table 7 needs to be communicated in a way that combines the different aspects of data in standardised diagrams, which leads to the visualisation and communication of results.

5.3 Visualisation and communication of results

Visualisation of results from the benefit-risk assessment plays a key role in the communication. The overall goal is to present as much information as possible in an easily comprehensible way. Four workshops were, as mentioned earlier, held and different visualisation approaches were tested. One of the first visualisation methods was a simple XY-plot, where overall risk and benefit scores were plotted, as seen in Figure 19, which shows that Drug 3 has a favourable benefit-risk balance, while Drugs 1 and 2 have the same balance, but where the overall benefit and risk score for Drug 2 is three-times the overall benefit and risk score of Drug 1. The next obvious question was: which drug is preferred if one is to choose between Drugs 1 and 2? This type of visualisation diminishes important information about the benefit and risk profile of drugs, causing a problematic situation for the decision makers.

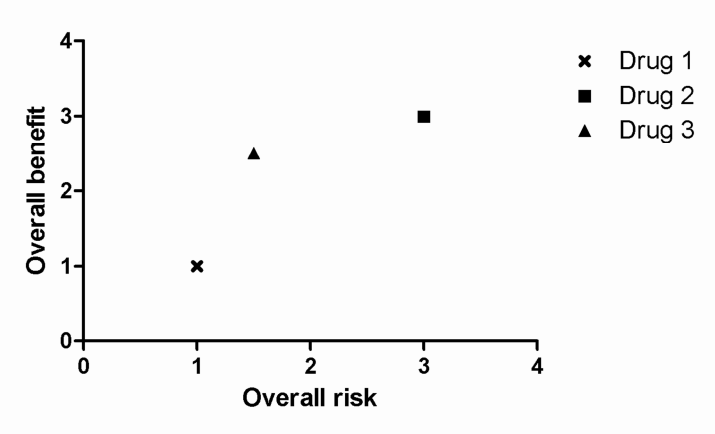


Figure 19: The Benefit vs. Risk balance for three drugs. Drug 3 has a positive benefit-risk ratio compared to Drugs 1 and 2, which have the same ratio but with different benefit and risk profiles.

There was a need to display all selected benefits and risks of drug and comparator, and two XY-plots, one for risk and one for benefits respectively, containing all the information were created and tested at an internal workshop in Novo Nordisk A/S (see Figure 20). In case of many criteria, people lost the overview and had problems at the attempt to summarise the results.

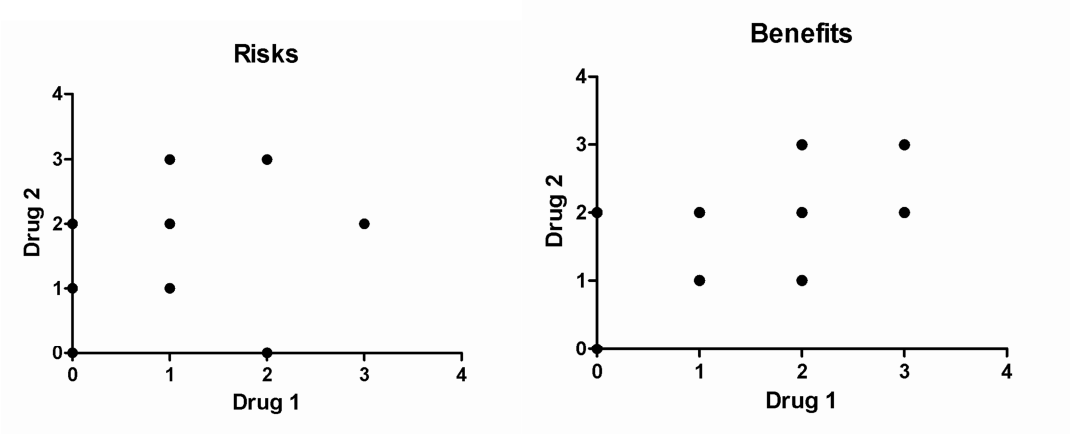


Figure 20: The weighted scores for both benefits and risk for two drugs are shown. Each dot represents a criterion.

The results of criteria sometimes ended on top of each other. This further hampered the assessment. Therefore, 3D scatter plots were created and presented at one workshop. Participants expressed that 3D scatter plots did not enhance their perception of data/results, and they used an unreasonable amount of time trying to interpret the plots.

Participants from one workshop requested plots that only focused on criteria where there was a difference between drug and comparator. This should reduce the unintended difficulties with the interpretation of XY and 3D plots. In Figure 21, the difference in weighted scores for benefits and risks for Drug 1 vs. Drug 2 are shown. However, the negative sign for Drug 2, indicating that it is superior, seemed to create confusion.

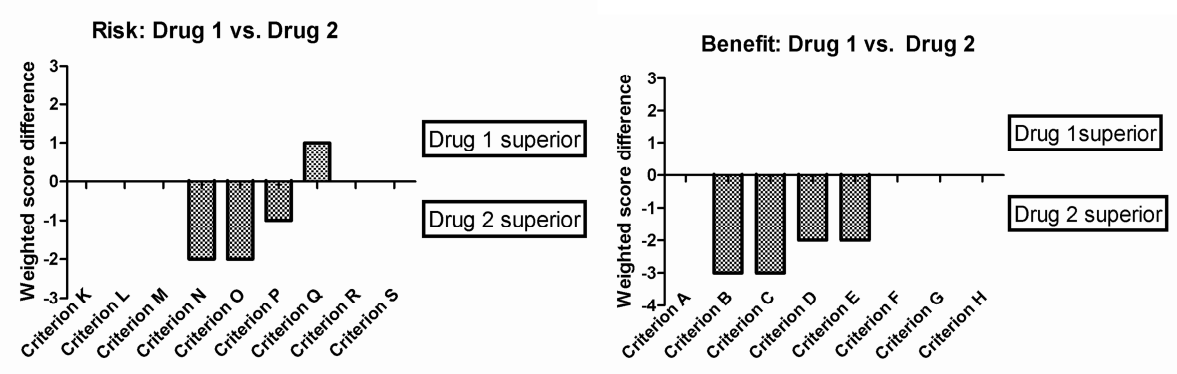


Figure 21: Shows plots for the difference in weighted scores for benefit and risk for Drug 1 vs. Drug 2. It is emphasised that another scoring scale of 1, 2 and 3 was used in this workshop, hence the scale for weighted scores is from -3 to +3.

In the following section, the focus will be on the visualisation of results from a single trial, multiple similar trials, multiple non-similar trials and dose-findings trials.

5.3.1 Presentation of weighted scores - Step 7

Based on the learning from the internal workshops at Novo Nordisk A/S, it was important to present the results in ways that enabled the human eye and mind to absorb as much information as possible in an intuitive and easily understandable approach. Furthermore, there is a need to address the important issue of correlations between criteria[5;9;16]. All the methods based on MCDA cannot avoid this step, and any attempt to ignore correlations can have negative consequences for the final result of the benefit-risk assessment.

Mussen et al.[16] clearly state that in a clinical benefit-risk assessment, using an MCDA approach, all criteria must be mutually preference independent, meaning that a score assigned to one criterion is independent of scores assigned to other criteria. Nonetheless, in a biological system, there is a high degree of correlation between the chemical and physiological processes at any given time, e.g. in a diabetes trial, fasting blood glucose is correlated to HbA1c, which is correlated to diabetic complications, etc. Thus, correlations cannot be avoided in a clinical and medical benefit-risk assessment, and therefore the MCDA value tree in its original form cannot be directly applied to medical decision making. Calculating overall benefit and risk scores is hereby not advisable.

To minimise the impact of correlations, the results of the assessment are visualised and communicated in tornado-like diagrams, as seen in Figure 22, where the *weight* of a criterion is depicted as the width of the box. The wider the box, the more important the criterion is. If the drug is superior, the colour green is used, yellow if there is no difference, and red is used if inferior.

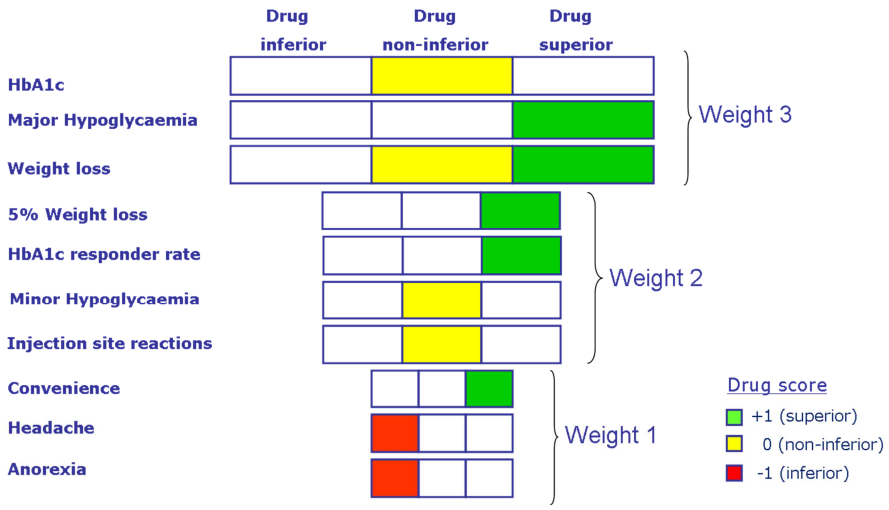


Figure 22: Results are presented in a tornado-like diagram. The width of the box indicates the relative importance of a criterion. The wider the box, the more important the criterion is. The colours indicate the results of the scoring. Green indicates superiority, yellow indicates non-inferiority and red indicates inferiority relative to a comparator**.**

5.3.1.1 Multiple similar trials

Similar trials can be combined to an overall benefit-risk assessment. Trials are assigned an impact factor based on their importance, e.g. the statistical power, key pivotal trial, etc. Once again, we use a simple scale of 1, 2 and 3 that reflect the uncertainty related to subjective assignment of values. Furthermore, the coarse scale increases transparency in the assessment compared to a much finer scale, which will undoubtedly give a wrong sense of accuracy in the assessment.

Scale:

1 = low impact

2 = medium impact

3 = high impact

A mean score for each criterion is calculated based on the weighted scores and the impact factors of the trials involved. If one trial has not measured a specific criterion, it should be made clear which trials the mean score is based on. In Table 8, an example is given. The mean is calculated to be 1/3; according to Table 9, the overall score is 0.

Table 8: The mean for criterion 1 is calculated to be 1/3, hence according to Table 9**,** the mean score is 0.

Trial	Impact factor	Criterion 1 score	Impact score of trial	Mean score for criterion
A	3	+1	+3	
B	2	0	0	
C	1	-1	-1	
Total	6		+2	2/6 = 1/3

Table 9: Scale for mean scores

	Drug inferior	Drug non-inferior	Drug superior
Mean	-1 to -0.5	-0.499 to 0.499	0.5 to 1
Score	Score -1	Score 0	Score +1

To gain an overall view of the performance of the drug compared to the comparator, a mean score is calculated for each criterion (any correlation is disregarded). Figure 22 shows three similar hypothetical trials - A, B and C - which are given the following impact factors: 3, 2 and 2. Figure 23 shows the individual results from three assessments, and figure 24 shows the overall assessment of the drug. The drug has an overall non-inferior to superior profile on all high and medium important benefit and risk criteria. For two criteria of low importance - headache and anorexia - the drug is inferior.

It is important to keep in mind that with overall results, the impact of correlations is higher and unpredictable. With this in mind, overall tornado diagrams can give an idea and direction of the overall performance of a drug.

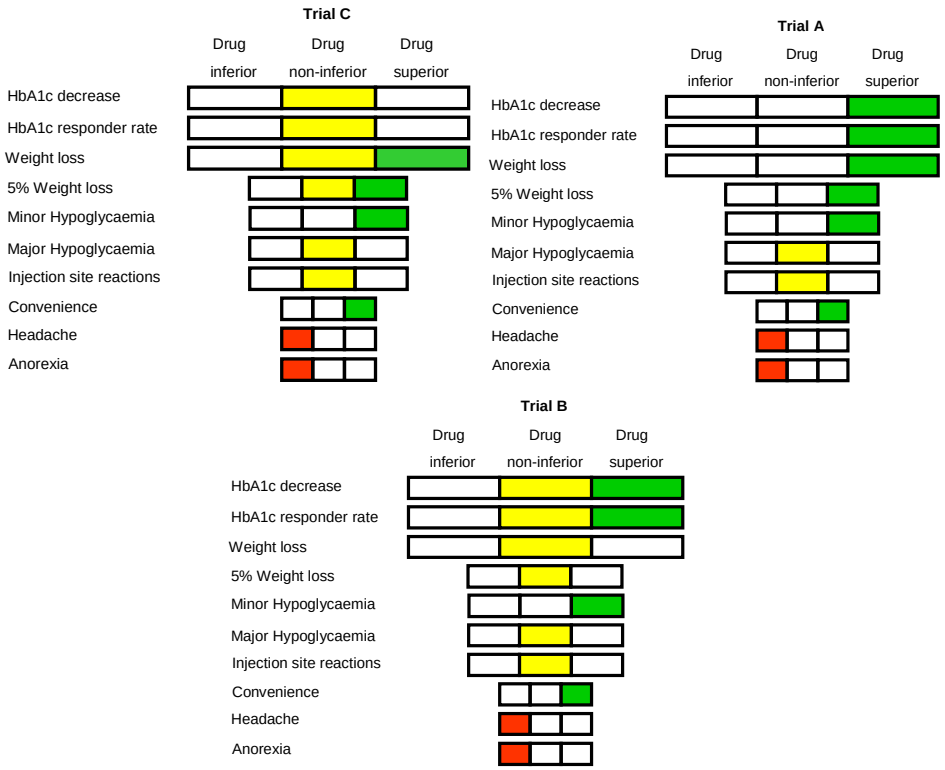


Figure 23: The result of the assessments for three hypothetical trials, A, B and C, with impact factors 3, 2 and 2, respectively.

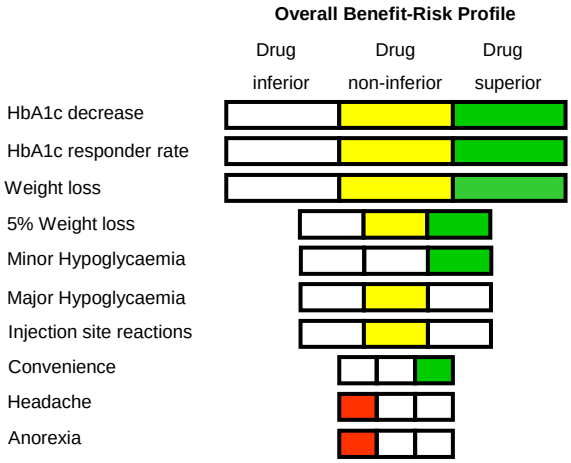


Figure 24: Based on the three individual benefit-risk assessments and the impact factors of each trial, mean scores are calculated for each criterion. The overall benefit-risk profile shows that the drug has an overall non-inferior to superior profile on all the high- and medium-importance benefit and risk criteria, while inferior on two criteria of low importance.

5.3.1.2 Non-similar trials

Combining non-similar trials to produce an overall benefit-risk profile is problematic, if not wrong, since different criteria are measured, and both patients and healthy volunteers are involved in the trials. However, to gain an overview of all the individual assessments, a new type of plot is necessary.

Some criteria in a given trial have more importance than criteria in another trial. Instead of assigning a common impact factor to the entire trial, as it was done under similar trials, individual impact factors are assigned to each criterion relative to all other criteria from all trials involved. In Figure 25a, a hypothetical phase 3 trial is shown, where individual impact factors are assigned to each criterion. In Figure 25b, the tornado diagram is converted to a column, where the weight and impact of each criterion are multiplied and converted to a colour nuance.

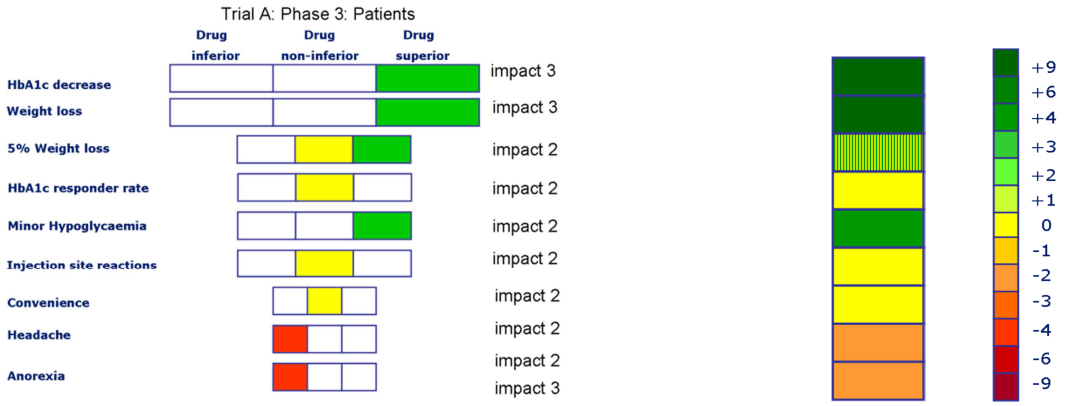


Figure 25: a) A hypothetical diabetes phase 3 trial with individual impact scores assigned to each criterion; b) the tornado diagram is transformed to a column, where the weight and impact of a criterion are visualised as a colour nuance.

Figure 26 shows how Trial A from Figure 25 is depicted alongside two other trials, B and C. In this hypothetical scenario, Trial B is a phase 2a trial with patients, mainly concentrating on safety and a few efficacy endpoints. Trial C is a phase 1 trial with healthy volunteers focusing on safety. Individual impact factors are assigned to each criterion across all trials, and impact-weighted scores are calculated. The scale is: -9,-6,-4,-3,-2,-1, 0, +1,+2,+3,+4,+6 and + 9. Not all criteria are measured in all trials, which contain different populations. However, in spite of diverging criteria, populations, etc., transforming these trials into columns based on their impact factors, gives an overview of what their similarities and differences are. Figure 26 shows that the drug is inferior on headache and anorexia across all trials. However, these two adverse events are weighted low in all trials and have a low impact. More interestingly, we see that the drug is superior on HbA1c, weight loss and minor hypoglycaemic events, which are all weighted high and which have a high impact on the results. The empty boxes simply indicate that a criterion is not measured in that trial, but applies to other trials.

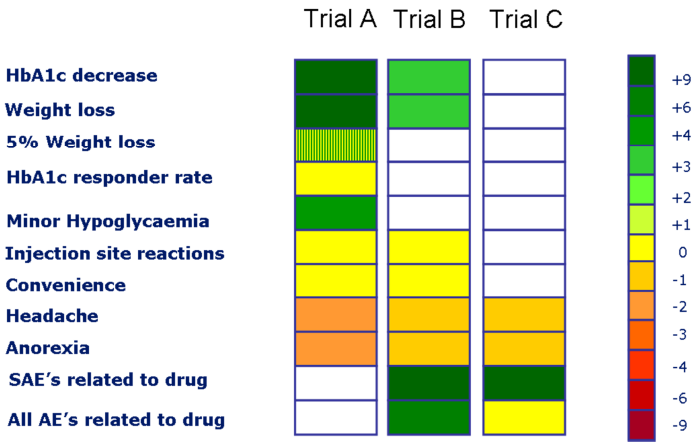


Figure 26: Each criterion in the three trials, A, B and C, are assigned individual impact factors. The impact-weighted scores are shown as colour nuances, stacked together and displayed. Trial A is a hypothetical phase 3 trial; trial B is a hypothetical phase 2 trial, and trial C is a hypothetical phase 1 trial.

5.3.1.3 Dose-finding studies

Some trials are dose-finding studies. For each dose, a tornado diagram is created. To enable optimal visual representation, each diagram is transformed into a column, as described above, where the weighted score of a criterion is visualised as a colour nuance, as seen above. The different doses are then stacked together and compared directly, as seen in figure 27. Based on the decision context, the optimal dose can be determined. If two doses are chosen, e.g. doses 3 and 4, an assessment can be performed, where one dose acts as a comparator. The differences between the doses are then captured in the resulting tornado-like diagram.

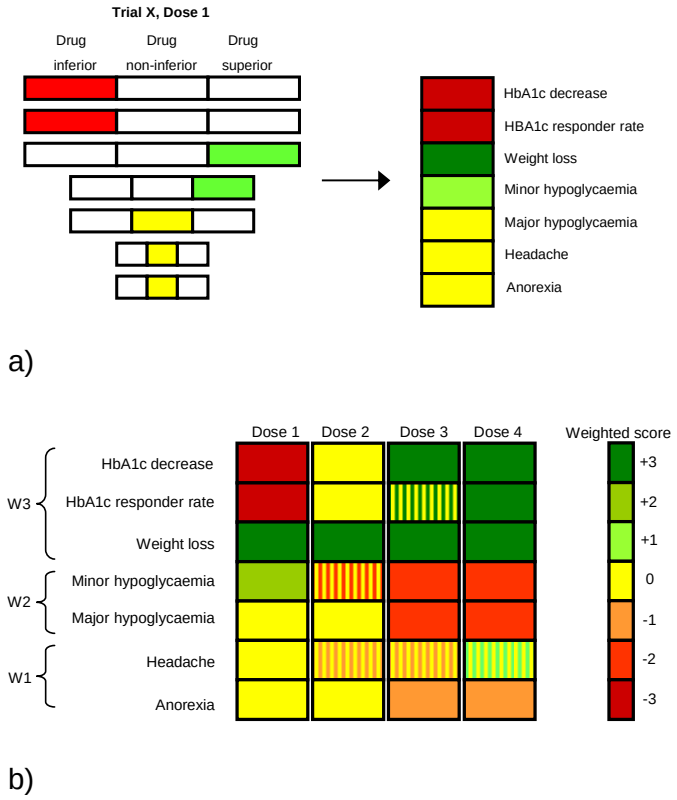


Figure 27: a) The diagram is transformed into a column. The weight of a given criterion is depicted as a colour scale. Interval scores are shown by shading lines. b) All doses are stacked together to enable optimal visualisation of the differences. Based on the decision context, the optimal dose can be estimated. W3 are high-importance criteria; W2 are medium-importance criteria and W1 are low-importance criteria.

5.3.2 Overall conclusions - Step 8

Steps 1 to 7 in the method provide the basis for step 8, the overall conclusion of the assessment. The method provides the basis for preservation of information and enables an overview of data in standardised diagrams. There are no attempts to calculate, e.g. a benefit-risk ratio, overall benefit and/or risk score. The method separates itself from decision theory-based methods on this aspect.

In the final step of the assessment, the most important benefits and risks are described explicitly in the therapeutic context, including a list of used benefit and risk criteria. The rationale and justification behind the selection of criteria and their weights are summarised, enabling transparency throughout the assessment and the final conclusion.

The definition of a clinically significant and relevant difference and the confidence level in event scoring are explained and justified. The observed effects, the expected acceptable

level of benefits relative to the risks, etc. are described. Uncertainties and the magnitude of evidence as well as their impact on the benefit-risk assessment are described. Any justification behind an interval score is provided, strengthening the transparency in the assessment and the overall conclusion. How benefits and risks may vary across different factors, such as age groups, sex, etc. is described. Issues that stand out, which need to be studied, are described. The expectations formulated in the decision context are evaluated. Finally, a recommendation and conclusion is given.

Every method must enable a sensitivity analysis. Otherwise, the method will be a black box and the results will be unreliable. A sensitivity analysis is performed to reveal any possible weak points in the assessment. First of all, weights are common for all drugs involved in the assessment. Therefore, any change in a given weight will simply move the weighted score proportionally with the change in weight. Secondly, the scores are data-driven and can therefore not be manipulated without changing the method itself. The ratio/cut-off point by which a drug has to perform better or worse in order to get a score different than zero can be varied to see the impact on the result. Any choice of ratio is justified, preserving the transparency in the method. Thirdly, an evaluation of the impact of uncertainty on the scores must be performed. Any interval score must be supported by justification and, when possible, by quantitative methods, preserving transparency in the assessment. The level of confidence in confidence interval scoring can be varied to see the effects on the objective scores.

5.4 Discussion and conclusion of Chapter 5

The method has a qualitative framework inspired by, e.g. MCDA, clinical significance, etc. The focus has been on the development of a transparent and structured framework, with focus on clinical significance and transparent visualisation of results. Subjective input is structured and justified and the clinical data itself determines the scores.

Clinical significance and relevance play an important role in the assessment of a drug or intervention, as emphasised by Boudes[25]. Commonly, clinical significance is based on personal experience and/or expertise, leading to discrepancies between different stakeholders, e.g. regulatory agencies and pharmaceutical companies. The definition of clinical significance is pivotal for the establishment of transparency in the benefit-risk assessment.

In the present context, the clinical significance of a drug is defined by the proportion of patients who experience a clear effect of a prescribed treatment with the drug. The drug is said to be clinically significant if the conditions are improved for, e.g. more than 2 out of 3 patients. This criterion differs from the criterion of statistical significance, as described in Chapter 4, of a positive drug effect, in that it measures the fractions of patients helped by the drug rather than demonstrating an improvement for the average patient. A very small difference in the effect can be statistically significant if the number of patients is high. Conversely, a significant effect of the drug can be statistically insignificant if the number of patients examined is small.

However, it is worth observing that even the smallest change is accounted for in the proposed methodology. To overcome this problem, one could simply define a cut-off point and only focus on the proportion of patients beyond this boundary, e.g. the proportion of patients experiencing an HbA1c below 7%.

For any disease area or treatment, key opinion leaders can define a ratio that represents a clinically significant and relevant difference. It is apparent that cancer will, compared to mild pain conditions, require a different ratio. The scoring ratio for confidence interval scoring can be changed in order to adjust the confidence level.

The lower the confidence level is, the higher the risk of a type 1 error will be (i.e. the risk that a null hypothesis is incorrectly rejected, when it is actually true). Here the idea may be, as described earlier, that the willingness to accept type 1 errors is accepted, if tendencies in the data can be captured, which otherwise would not have been captured with a normal 95% confidence interval. The tendencies can then undergo careful analysis and discussions, and

probable or possible related events to the treatment with the drug can be investigated in future studies.

A comparator is not always available. Under these circumstances, placebo treatment may be used as the comparator or, if this is not practical, the untreated disease itself may provide baseline values. For the efficacy parameters, usually biomarkers or other biophysical/biochemical parameters, the baseline values can be used to create the difference distributions once the end-of-trial values for the drug have been obtained. For the risk parameters, the background incidences for the untreated disease can be used as baseline parameters, both in the cases where new risks are introduced by the drug and in cases where the considered types of risk already exist in the untreated state, but are promoted by the drug.

Uncertainty and marginal scores are handled by the qualitative features of the assessment methodology. However, quantitative methods can also be drawn to support the assessment of different parameters. Non-parametric re-sampling can be used to build multiple trials from a given dataset. For each such trial, scoring of a given criterion can hereafter be performed, and the resulting distribution of scores can be used to decide, for instance, whether an interval score is more appropriate than a point score. A specific problem arises in the re-sampling of events. If no events occur in a study population, e.g. 0 events of major hypoglycaemia in a study population of 50, regardless of how many times the re-sampling is performed, no events will occur. That is a limitation of non-parametrical re-sampling.

Parametric re-sampling represents another possibility for estimation of uncertainty in special cases. Such methods have not yet been applied in connection with the present assessment approach, but the general idea is that the assessments should be supported by statistical methods when required.

I believe that a qualitative evaluation of uncertainty and evidence is the pivotal step. Evaluation of uncertainty is not only related to within-trial uncertainty and marginal scores. A particular problem arises from the substantial differences that exist in the amount of supportive evidence between a new drug and an already-established drug on the market. To

overcome this difficulty, a particularly critical evaluation of the existing evidence will be required, and uncertainty around an objective score for the drug candidate or overwhelming positive evidence for the existing drug may lead to adjustment of the scores, based on qualitative judgement and justification.

A flat structure is chosen because there is no interest in calculating a benefit-risk ratio. Instead, it is wished to minimise the impact of correlations between criteria, to avoid linear combinations of criteria and to preserve as much information as possible. The presented structure has proved very suitable in coping with these issues, and it offers transparent communication of the results. Major issues, like serious adverse events, are handled more carefully in tornado-like diagrams, where no information is lost. The tornado-like diagram depicts all the differences between a drug and comparator and can facilitate a discussion of essential issues, e.g. differences in serious adverse event profiles between drug and comparator. This means that if the number of subjects is large enough, a distribution of the individual measurements from each subject can be calculated numerically for both drug and comparator. Using these two empirical distributions, the empirical difference distribution can be calculated.

Difference distributions are used for scoring, and in this thesis data is fitted to distributions, e.g. normal distribution, but the scoring is not restricted to this. Difference distributions should be created empirically when data allows it.

In classical statistics, the aim is to assess whether a given difference for a given criterion is a matter of uncertainty or systematic behaviour. A very small difference can be statistically significant if the number of subjects in the trial is high, and a very large difference can be discarded as non-significant if the number of subjects is low. The fundamental behind the method based on proportions is to compare fractions of data. The method is independent of the number examined, and differences seen in small and large clinical trials will consequently reach the same score. The difference distributions create the basis for a qualitative discussion based on quantitative results.

6 Learning and results

In this section, the concrete different learning and results gained through the PhD project will be discussed and presented.

The proposed method was developed as an iterative process, where a simple version was firstly created and then tested in a pilot study at a workshop in Novo Nordisk A/S. The method was then modified based on learnings from the workshop, and was tested again in a second pilot study in another workshop. In total, four workshops and pilot studies were thusly held in Novo Nordisk A/S. One of the pilot tests involved a marketed product by Novo Nordisk A/S. The remaining pilot tests involved three drugs, which are still under development.

The method was presented to the Danish Medicines Agency and several medical officers were interviewed afterwards. A summary from these discussions will be presented.

In collaboration with Bispebjerg Hospital, Department of Clinical Pharmacology, the method was tested on a schizophrenia database.

The method has also been tested on Ketek®, in cooperation with Dr Christine E. Hallgreen and PROTECT. Data were obtained from EPARs from the European Medicines Agency's (EMA) website.

To illustrate the practical use of the method, I will present the results gained from the collaboration with the Copenhagen University Hospital (Rigshospitalet), Department of Clinical Pharmacology, where the method was used on a colorectal cancer database. Furthermore, the results of the application of the method to a project concerning schizophrenia will also be presented. Finally, the result gained from the collaboration with PROTECT WP5 on telithromycin (Ketek®) will also be presented.

6.1 The four workshops

Four different drugs were evaluated in pilot tests. Each drug brought about new issues and new data types that were not seen with the previous drug. The first two workshops were 1-

day exercises, and the last two workshops were 2-day exercises. However, because of issues of confidentiality, I am not allowed to show data and results from these pilot studies and workshops due to issues of confidentiality. Instead, I will elaborate on alternative results and experience from these workshops. First of all, the first two workshops indicated that one day was insufficient time to have in-depth discussions about which criteria to choose, their weights, the definition of clinical significance, etc. The participants often needed some thorough explanation of concepts like “weighting” and “scoring”, which are unknown terms to people without knowledge of decision theory.

Entire project teams were invited to all four workshops, that is to say the project manager, the international medical director, the regulatory affairs associate, the statisticians, the medical writer, the safety surveillance advisor, the clinical pharmacologist and the medical affairs advisor. The project team decided beforehand which trials an assessment should be conducted for.

In the first two workshops, the definition of decision context, decision profile, weighting and a definition of clinical significance were the features of an around-the-table discussion. Within a few minutes, the most dominant person in the group, typically a leader, led the discussions from start to end, and the rest of the group more or less accepted this. As a consequence, the discussions were one-sided and superficial. It was almost as if the official company opinion was being presented at a press conference.

To avoid this scenario in the last two workshops, the group defined the decision context in consensus, but the selection of criteria, weighting and justification was first performed individually. Every group member then presented his/her results and once all members had presented their results, the group decided on which criteria should be chosen in the decision profile and their respective weights. In this way, every group member had a chance to express his/her opinion. The discussions were more lively and differentiated, and resulted in an in-depth undertaking of all issues from different points of view.

Another lesson learnt from these workshops was that justification of choices increased transparency and credibility of the assessment. Once a preference is argued and justified, it

gains credibility. Every preference had to be justified, that is to say the definition of clinical significance, the criteria to be included in the assessment, the weights of criteria, uncertainty evaluation, etc. In this way, the assessment could be a dynamic entity that can be changed and justified as new knowledge, and information is gained through its life cycle.

In general, all four project teams participating in the workshops expressed that the structured approach to the benefit-risk assessment was highly valuable, enabling a thorough discussion of all aspects of data and the choices related to the final decision.

6.2 Discussions with the Danish Medicines Agency (DMA)

The benefit-risk assessment method was presented to the Licensing Department at the Danish Medicines Agency (DMA), and several medical officers were later interviewed. The aim was to clarify their views on benefit-risk assessment in general and on the current method. Three senior medical officers and one chief medical officer were interviewed individually. The list of questions can be seen in Appendix A. Three topics were discussed:

1. Use of methods/models in DMA.
2. How can benefit-risk assessments be improved?
3. The current benefit-risk assessment method.

Use of methods/models in DMA

The medical assessors use no specific model or method in their evaluations and assessments. In the evaluation of a new drug application (NDA), the Day 80 benefit-risk assessment of a drug (Day 80 AR, EMA) and the following are, to a large extent, used as the basis for the assessment by the assessors: CHMP guidelines, guidelines from international clinical societies, ICH guidelines, epidemiological knowledge and personal experience and expertise. Controversies between the rapporteur and co-rapporteur occur seldom. When a

difference of opinion arises, it is usually based on differences in personal experience and expertise.

Currently, there is no validated structured weighting of benefits and risk on a value scale. The final benefit-risk assessment is based on the structured approach in the D80 AR with thorough discussion of the balance between beneficial effects and the uncertainties related to them vs. the unfavourable effects and their uncertainties.

How can benefit-risk assessments be improved?

It was clearly highlighted by the medical officers that the existing quantitative data-driven tools/methods for benefit-risk assessments would be beneficial, provided that these methods were adequately validated.

It was generally agreed upon that the industry could enhance the quality of their own benefit-risk assessments by the use of structured approaches. This would enhance transparency in their benefit-risk assessments and overall conclusions on their drug.

Transparency could be enhanced by the implementation of quantitative data-driven methods for the weighting and assessment of data. It was expressed that the agencies and the industry, although looking at the same data, can have differing interpretations of data. In this context, a structured approach, where different types of clinical parameters are weighted and justified, would be desirable.

The current benefit-risk assessment method

In general, it was stated that if the method is adequately validated, the method has the potential to be used in the final steps of drug approval, and there was a general positive impression of the visualisation of results; the principle of objective scoring was seen as an interesting aspect. When validated it could increase transparency.

The following more specific statements were made: it was expressed that although there are no obvious pitfalls in the method, these can only be detected once the method is used and validated in a real-life setting. We were therefore encouraged to use the method, and the

use of the data-driven tools was especially emphasised. The weighting of benefits and risks and the eight-step successive approach was seen as an opportunity to perform more transparent assessments.

Once validated, the method could be a good supplement to the assessments already made by the industry. The visualisation of events in scoring tables and the difference distributions can be useful tools in detecting trends and communicating results.

6.2.1 Conclusions of the interaction with DMA

Validation is a term that seems to haunt the development of benefit-risk assessment methods. Both internally in Novo Nordisk and externally in DMA, there is focus on validation. However, results from such tools cannot be validated, since it is problematic to validate human judgement, which instead can be justified more or less convincingly.

In the proposed method, there are several steps with subjective input, and the results of the assessment are consequently influenced by these steps. A sensitivity analysis, as described under step 8 in the method, can be performed, but this is not to be confused with validation. The term validation is typically associated with statistical methods and is wrongfully transferred to the field of benefit-risk assessment. Finally, the value of the method can only be judged by the use and acceptance it can gain from the regulatory agencies and from other pharmaceutical companies.

6.3 Colorectal case

The present study tests the proposed methodology, which can be used in drug development as well as in data-driven assessment in the comparison of two (or more) options. The purpose of the present section is to demonstrate the method by analysing the obtained results by Afzal et al.[117] in a study of colorectal cancer treatment. It will be demonstrated how these data can be weighted, scored and communicated in a structured and transparent

manner with a focus on the quantification of the clinical significance and on the general tendencies in data[118].

6.3.1 Colorectal cancer and 5-Fluoruracil

In 2008, colorectal cancer had a worldwide incidence of 1,233,700 cases and 608,700 cases of estimated deaths. Together with lung, breast and prostate cancer, colorectal cancer has one of the highest incidence and death rates[119]. The work of Afzal and co-workers focuses on 5-Fluorouracil (5-FU), which is widely used to treat solid tumours, including colorectal cancers. The cytotoxicity of 5-FU depends primarily on two active metabolites: 1) Fluorodeoxyuridine monophosphate (5-FdUMP) inhibits the thymidylate synthase (TYMS) enzyme, and 2) Fluorouridine triphosphate (5-FUTP) impairs RNA function and thereby induces cell toxicity. The inhibition of the TYMS enzyme is dependent on and enhanced by intracellular 5,10-methylenetetrahydrofolate[117].

Increased sensitivity of cancer cell lines to 5-FU is correlated with decreased expression or activity of dihydropyrimidine dehydrogenase (DPYD), methylenetetrahydrofolate reductase (MTHFR) and TYMS as well as increased activity or expression of oratate hosporibosyltransferase (OPRT or UMPS). Studies investigating the association of TYMS, DPYD, OPRT and MTHFR polymorphisms or expression with survival in adjuvant 5-FU-based treatment of colorectal cancer have yielded contradictory results, especially regarding TYMS and MTHFR[117].

Most studies investigate the association of individual polymorphisms with disease-free survival (DFS) or overall survival (OS). Theoretically, the 5-FU metabolic phenotype is better explained by multi-gene and pathway-oriented analysis rather than single gene analysis. Systematic multi-polymorphism combinations, by the use of the multifactor dimensionality reduction method (MDR), have revealed that low expression alleles in thymidylate synthase (*TYMS*) are associated with decreased DFS and OS. A specific MDR-derived combination

of DPYP and TYMS VNTR polymorphisms was observed to be associated with increased DFS [132]. Figure 28 shows an overview of the metabolism 5-FU.

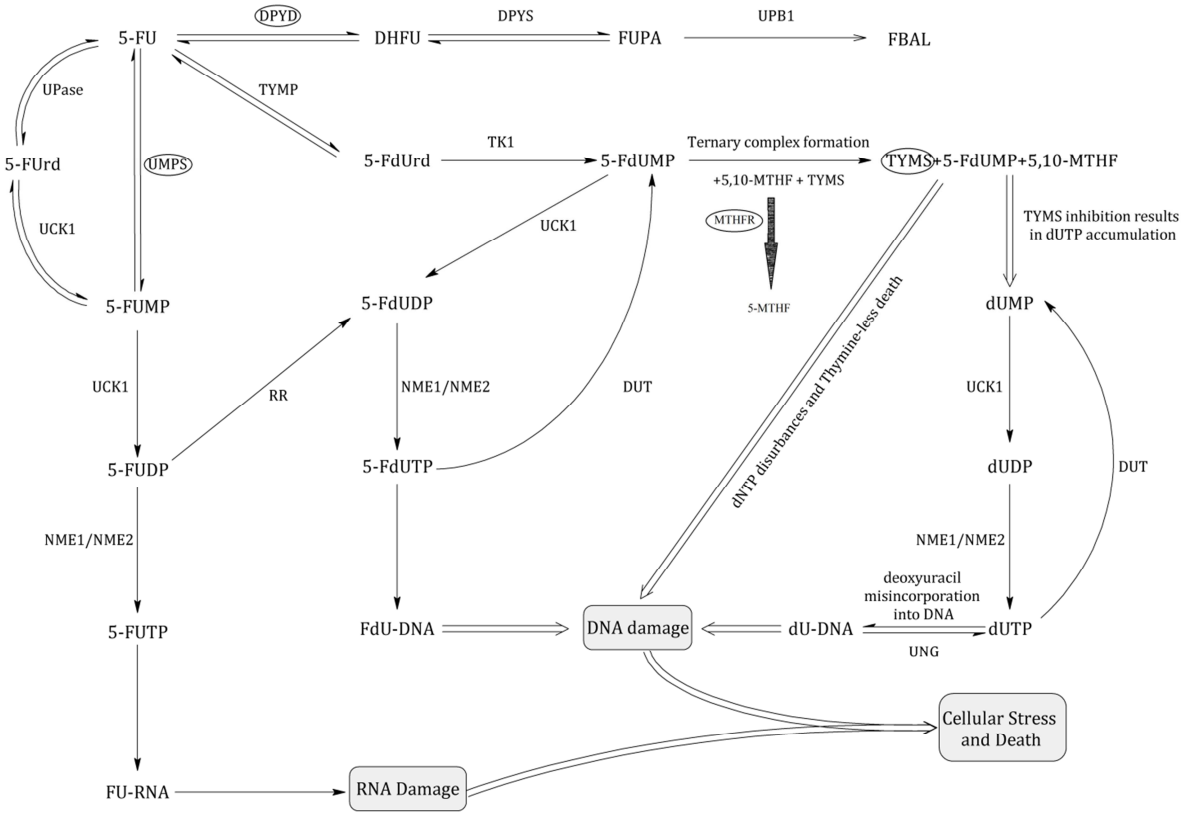


Figure 28: Overview of 5-FU metabolism. The 4 central enzymes that Afzal and co-workers focus on are encircled in the figure: dihydropyrimidine dehydrogenase (DPYD), methylenetetrahydrofolate reductase (MTHFR), thymidylate synthase (TYMS) and orotate phosphoribosyl transferase (OPRT/UMPS).
Source: Afzal S, Gusella M, Jensen SA, Vainer B, Vogel U, Andersen JT, et al. The association of polymorphisms in 5-fluorouracil metabolism genes with outcome in adjuvant treatment of colorectal cancer. *Pharmacogenomics*. 2011 Sep;12(9):1257-67.

6.3.2 Material

The study by Afzal et al. analysed data from 302 patients, where ten polymorphisms in genes involved in 5-FU pharmacodynamics and pharmacokinetics had been studied. Multifactor dimensionality reduction (MDR), a non-parametric method, was used to identify genetic interaction profiles associated with outcome. An MDR analysis of all the polymorphisms and the functional classifications of *MTHFR*, *TYMS* and *DPYD* has previously shown that variant alleles in *DPYD* and the *TYMS* VNTR polymorphism are associated with improved DFS[117].

Median follow-up time was 5 years (up to 11 years). This is probably related to the fact that many patients surviving more than 5 years were seen as cured. The last follow-up date was 30 August 2007. All patients were Caucasian with Dukes' stage B2 and C treated at Rigshospitalet (Copenhagen University Hospital) with surgery and the Mayo regimen (Levofolinate 10 mg/m^2, 5-FU 425 mg/m^2/day, days 1–5, every 28 days, six cycles). DNA was isolated from formalin-fixed paraffin-embedded (FFPE) tumour tissue, with maximum 50 % normal tissue. Clinical data and tumour pathology were reviewed retrospectively.

Table 10: Clinical data from the studied cohort.

	N=302
Age at diagnosis (median years, range)	61 (19-85)
Sex	
Male	151 (50 %)
Female	151 (50 %)
Median follow-up (years)	5.3 (0.1 – 11.3)
DFS	
Events	142 (47 %)
Censored	160 (53 %)
OS	
Events	127 (42 %)
Censored	175 (58%)
Stage	
B	37 (12 %)
C	265 (88 %)
Tumour grade	
1	91 (30 %)
2	129 (43 %)
3	82 (26 %)
Missing	2 (1 %)
Tumour site	
Colon	246 (81 %)
Rectum	56 (19 %)

Table 11 shows the 269 patients that were eligible for this study, distributed according to their MDR classification. The relatively high number of eligible patients indicates that the

association of *DPYD* and *TYMS VNTR* is highly represented amongst patients and that the classification of patients into two groups according to their MDR classification could have significant clinical implications for patients. The MDR-1 group consists of patients with the combination of variant alleles in the *DPYD* gene and the *TYMS VNTR* polymorphism, selected by the MDR algorithm as being associated with improved DFS.

Table 11: The patient population divided by the MDR classification

	Number of patients (N= 269)
MDR 1	111
MDR 0	158
Missing	33*

*Sufficient DNA material could not be obtained for 33 patients.

6.3.3 Method

The data are assessed by the benefit-risk assessment method, as described in Chapter 5. In the following case, it will be demonstrated that not only is the proposed methodology suited for benefit-risk assessment in drug development, it is also suitable for data-driven assessment of any given two options. In this case, the two options are given by the combination of genetic polymorphisms. Figure 29 recapitulates the eight successive steps comprising the qualitative framework.

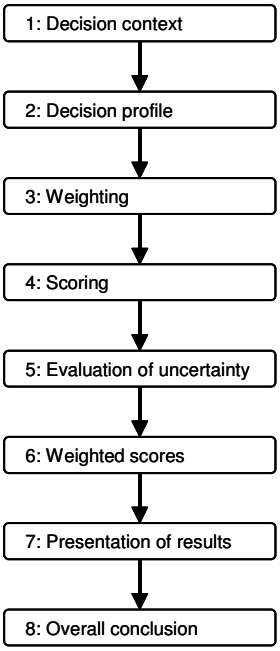


Figure 29: The eight successive steps in the benefit-risk assessment method.

6.3.4 Results

1 Decision context: The most interesting question is: how well do two groups of patients with colorectal cancer, but different genetics, according to the MDR classification, respond to the same treatment? To answer this question, a head-by-head comparison on the basis of cure rate, survival rate, time-to-death (TTD), time-to-relapse (TTR), and main adverse drug reactions is performed.

As discussed in Chapter 5, clinical significance must be predefined before any data are revealed. In the present context, clinical significance is defined as 12 out of 20 patients in one of the two groups having to show a difference for continuous data, e.g. time-to-death and time-to-relapse relative, in order for the difference to be considered clinically significant and relevant. The choice of the ratio is based on the severity of the disease, where even a small difference can have a clinical impact.

In the assessment of event criteria, the level of confidence is lowered to 2/3 (= 66.7%). However, for cure and survival rate, the confidence level is lowered to 60% in accordance with the argumentation used for TTD and TTR. A low confidence level is chosen to capture

tendencies in data that otherwise would not have been captured with the conventional level of confidence of 95%. A lower confidence level is associated with higher risk of a type 1 error (i.e. the risk that a null hypothesis is incorrectly rejected, when it is actually true).

This means that any tendency captured must be evaluated thoroughly in order to determine whether there is any plausibility to it. The data that are the basis of this study, as described earlier, comprised 302 patients with colorectal cancer treated at Copenhagen University Hospital.

2 Decision profile: Based on the decision context, the following criteria are chosen: cure rate, survival rate, TTD, TTR, infections, bleedings, mucositis/stomatitis, nausea/vomiting, hand-foot skin syndrome, diarrhoea, arthralgia/myalgia, myocardial ischemia and fatigue, as seen in Table 12. These were also the main criteria that were available in the database from Copenhagen University Hospital. The selected criteria are justified to represent the most relevant aspects in the evaluation of differences between options when assessing colorectal cancer.

3 Weighting: As described in Chapter 5, criteria are weighted on a value scale to enable comparison. The weights are based on the relative importance of a difference between two options. Each criterion is assigned a weight/importance of 1 (low), 2 (medium) or 3 (high). The weights are comparable across all criteria in the assessment, meaning that two criteria with the same weight are equally important. Weighting is independent of the datasets. All weights must be justified for documentation and communication purposes. Different stakeholders (e.g. clinicians, patients, pharmaceutical industry, health authorities, etc.) may have different opinions about the weights.

In this study, weighting is performed from a clinical point of view. TTR and TTD are primary endpoints, since both patients and clinicians are interested in survival or cure. These two criteria are therefore of high importance, meaning that a difference between the two MDR groups will have clinical implications.

Criteria such as infections, myocardial ischemia, bleedings, mucositis/stomatitis, hand-foot skin syndrome and diarrhoea are all considered of medium importance, meaning that a difference between two options will probably have clinical implications, e.g. if the performance for the high-importance criteria is equal between the two options. The medium-importance criteria are often difficult to treat and can have consequences for the patients, which are occasionally very serious or fatal.

The low-importance criteria, arthralgia/myalgia, fatigue and nausea/vomiting can often be treated, e.g. pharmacologically. Although these events can be very severe, they are rarely life threatening.

Table 12: Shows the selected criteria and their weights

Criterion	Weight
Cure rate	3
Survival rate	3
TTD	3
TTR	3
Infection	2
Myocardial ischemia	2
Bleeding	2
Mucositis/stomatitis	2
Hand-foot skin syndrome	2
Diarrhoea	2
Arthralgia/myalgia	1
Fatigue	1
Nausea/vomiting	1

4 Scoring: Based on the MDR classification, the TTR and TTD curves are created for the MDR-0 and MDR-1 groups, as seen in Figure 30. There is no follow-up after 8 years, where a considerable number of patients are still alive. We need to know the entire distribution to be able to calculate the difference distribution between two distributions[112]. The Kaplan-Meier curves are therefore fitted to Weibull distributions that are often used to describe and analyse survival data. This distribution seems to also offer a reasonable description of our data.

Figure 30a shows the TTD as a function of MDR-1 (red) and MDR-0 (blue) respectively. As seen in Figure 30a, a substantial number of patients survive more than 8 years in each group, and the dashed lines represent the “survival rate” of colorectal cancer for the two MDR classifications, where 67/111 (60%) patients in the MDR-1 group survive compared to 52/158 (33%) patients in the MDR-0 group. The curves are normalised (data not shown) to represent time-to-death for all patients dying during the study. The difference distribution in Figure 30b shows that regarding the patients dying during the study period, 64% in the MDR-1 group have a longer time-to-death relative to the patients dying in the MDR-0 group. Figure 30c-d shows the TTR and is to be interpreted in the same manner. The dashed line represents the “cure rate”, where 66/111 (59%) patients in the MDR-1 group are cured compared to 62/158 (39%) patients in the MDR-0 group. Figure 30d shows that 55% of the patients surviving in the MDR-1 group have longer time-to-relapse relative to the patients surviving in the MDR-0 group.

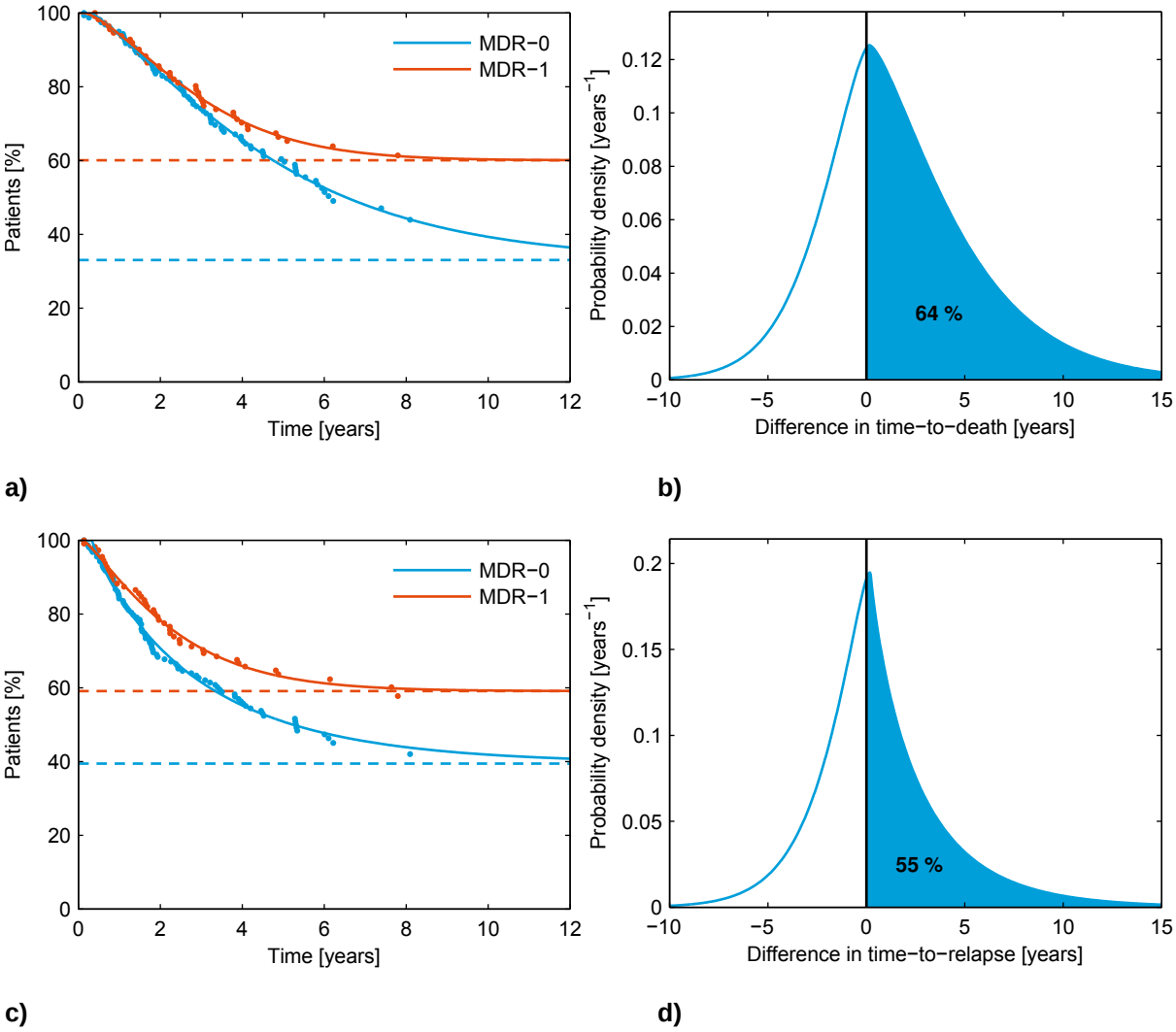


Figure 30: a) Time-to-death (TTD) curves for the MDR-0 and MDR-1 groups. The dashed lines show the level of survival. b) Difference distribution for TTD, c) Time-to-relapse (TTR) curves, and the dashed lines show the level of cure. d) Difference distribution for TTR. 67/111 (60%) patients in the MDR-1 group survive compared to 52/158 (33%) patients in the MDR-0 group, while 66/111 (59%) patients in the MDR-1 group are cured compared to 62/158 (39%) patients in the MDR-0 group.

The dashed lines in Figure 30a and c represent and estimate the “true” level of survival rate and the “true” cure rate. In Table 13, the number of patients that will either survive the disease or be cured from it is shown.

Toxicity was recorded and graded according to the Common Toxicity Criteria (CTC) (National Cancer Institute, Common Toxicity Criteria version 2.0)[117]. Data was processed further so that toxicity was dichotomously classified into none-to-moderate (grade 0–2) and severe (grade 3–4) toxicity for the MDR classification. It was decided to focus on differences

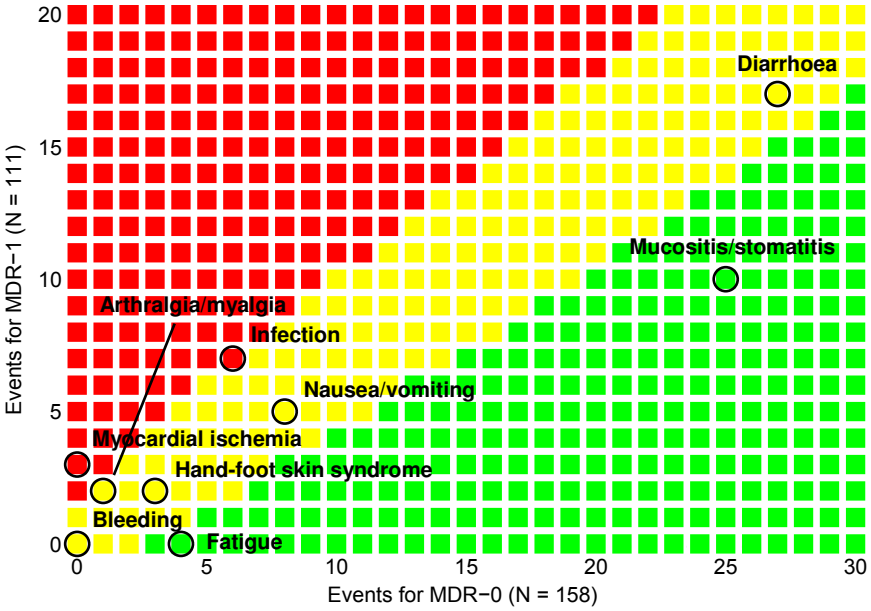
in the severity of adverse drug reactions between the two MDR classifications. Two one-sided binomial tests are performed by the use of the Clopper-Pearsons exact method[115].

Table 13: Number of grade 3-4 toxicity of adjuvant chemotherapy based on the MDR classification, and number of "true" cured and survivors.

Criterion	MDR-1 (N=111)	MDR-0 (N=158)
Cured	66	62
Survivors	67	52
Infections	7	6
Bleedings	0	0
Mucositis/Stomatitis	10	25
Nausea/Vomiting	5	8
Hand-foot skin syndrome	2	3
Diarrhoea	17	27
Arthralgia/Myalgia	2	1
Myocardial ischemia	3	0
Fatigue	0	4

Data from Table 13 are visualised in the scoring tables shown in Figure 31, which shows that patients in the MDR-1 group have a superior profile with regard to mucositis/stomatitis and fatigue. For myocardial ischemia, the MDR-1 group has an inferior profile with 3 cases of myocardial ischemia versus 0 cases in MDR-0. The MDR-1 group is also inferior on the infection criterion. For the remaining criteria, the MDR-1 group is non-inferior. Figure 31b shows that patients in the MDR-1 group have a clear superior profile with regard to cure and survival.

a)



b)

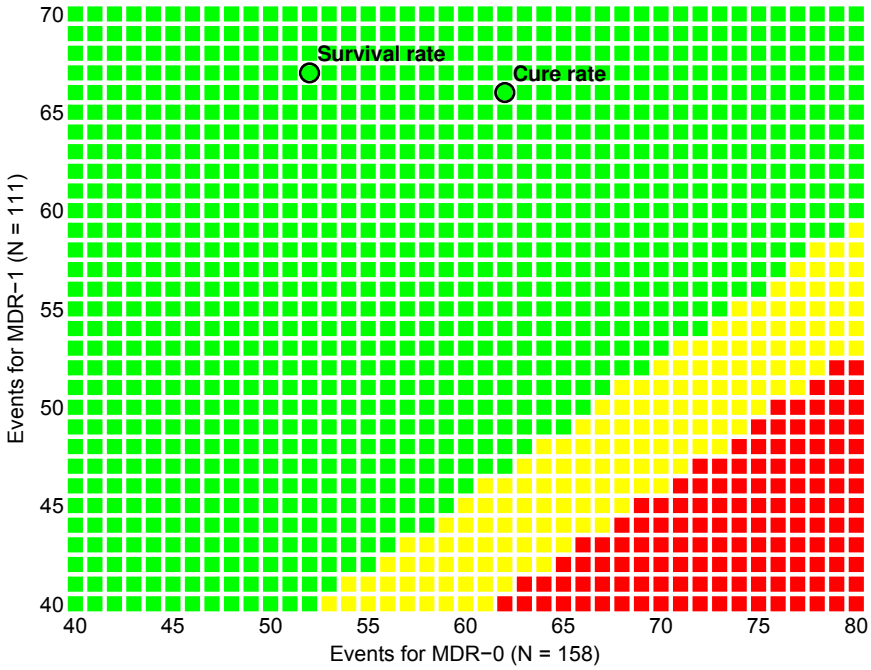


Figure 31: Data-driven scoring based on confidence interval scoring. a) Scoring table showing that MDR-1 is inferior on myocardial ischemia and infections, while superior for mucositis/stomatitis and fatigue. b) Scoring table for number of survivors and cured patients, based on the estimates from Figure 30. MDR-1 shows a superior profile with regard to both cure and survival rate.

5 Evidence and uncertainty: The objective scores must be reviewed critically. Uncertainty related to the data-driven scores can be estimated quantitatively on data level, e.g. by non-parametric or parametrical resampling techniques. However, it is believed that a qualitative

evaluation of uncertainty and evidence is more decisive and should be the backbone of an evaluation and only supported by quantitative measures when necessary.

The objective scores are based on a total of 269 patients, distributed in the two groups. The basis for the scores is relatively weak, due to the low number of patients, and this is accounted for in the data-driven scoring method in a simple way: The objective scores of borderline criteria are changed to intervals. Arthralgia/myalgia, infections and fatigue are all borderline and the scores are consequently changed to intervals, as seen in Table 14. Time-to-death curves show that 64% of patients dying in the MDR-1 group have a longer time-to-death, while the difference distribution for time-to-relapse shows that 55% of the patients surviving in the MDR-1 group have longer time-to-relapse relative to the surviving patients in the MDR-0 group. According to the definition of clinical significance, the MDR-1 profile is not superior to the MDR-0 profile with regard to time-to-relapse, but there is a tendency in favour of MDR-1. However, this difference is regarded as clinically not significant.

For the criteria infections, arthralgia/myalgia and fatigue, only one more case in one of the groups will shift the score. Based on these observations, the scores for these criteria are changed to intervals.

6 Weighted scores: The weights and scores are multiplied to produce weighted scores that enable the direct comparison of different criteria between the two options. The weighted scores capture both the importance and performance. Table 14 shows criteria, their weights, scores and weighted scores.

Table 14: Shows criteria, weights, scores and weighted scores in the MDR-1 group.

Criterion	Weight	Scores	Weighted Scores
Cure rate	3	1	3
Survival rate	3	1	3
TTD	3	1	3
TTR	3	0	0
Infection	2	-1→0	-2→0
Myocardial ischemia	2	-2	-2
Bleeding	2	0	0
Mucositis/stomatitis	2	1	2
Hand-foot skin syndrome	2	0	0
Diarrhoea	2	0	0
Arthralgia/myalgia	1	-1→0	-1→0
Fatigue	1	0→1	0→1
Nausea/vomiting	1	0	0

7 Visualisation: The results of the assessment are visualised and communicated in a tornado-like diagram, as seen in Figure 32, where the *weight* of a criterion is depicted as the width of the box and the *score* is represented by the colour. The wider the box, the more important the criterion is. If the MDR-1 is superior, the colour green is used, yellow if there is no difference, and red is used if the MDR-1 is inferior. Figure 32 shows that patients in the MDR-1 group have an equivalent or better outcome for the most important criteria, cure rate, survival rate, TTD and TTR, while being only inferior to non-inferior for one medium-importance and low-importance criterion - infections and arthralgia/myalgia respectively. In the tornado diagram, both of these are marked as inferior (red). Patients in the MDR-1 group have a clearly inferior profile for only one medium-importance criterion - myocardial ischemia.

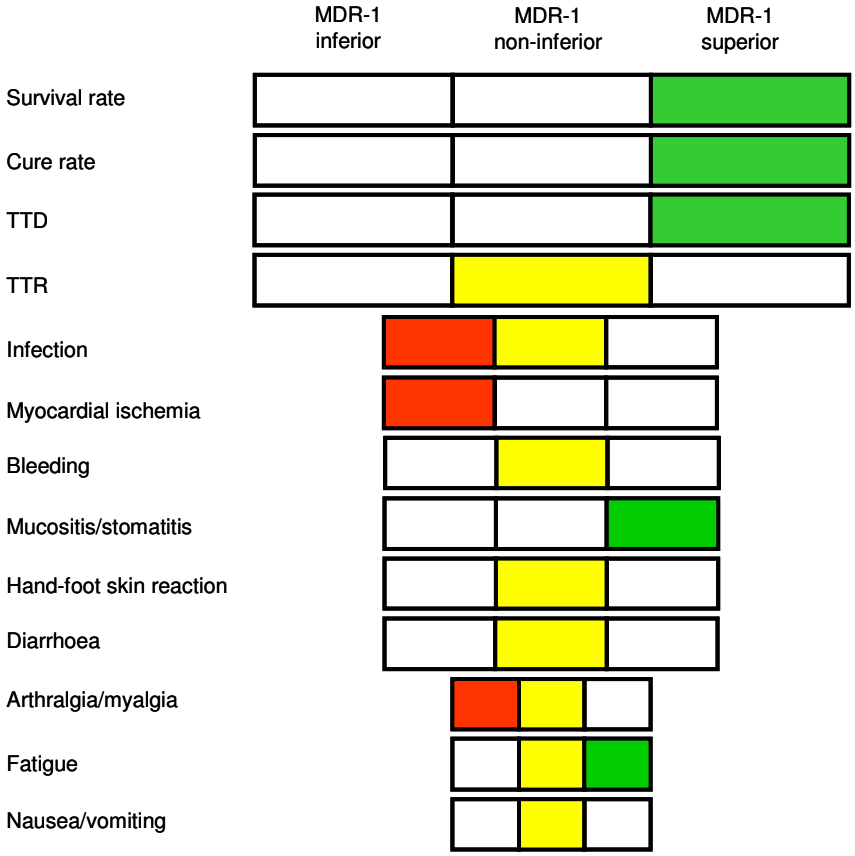


Figure 32: The results of the assessment are shown, where patients in the MDR-1 group have a superior profile with regard to three out of four most important criteria, survival rate, cure rate, and TTD, while non-inferior for TTR.

8 Overall conclusion of the assessment: A clinically significant and relevant difference for the high-importance criteria cure rate, survival rate, and TTD was found in favour of the MDR-1 group, while for TTR the results were non-inferior, but a tendency in favour of the MDR-1 group is seen. A higher risk for the medium-importance criterion myocardial ischemia and a slightly higher risk for the medium-importance criterion infection were seen in the MDR-1 group, which needs to be studied. Because of the limited number of patients, the results for the criteria TTR, infections, arthralgia/myalgia and fatigue are not conclusive and are only seen as trends/tendencies.

The clinical implications of this study are that the genetic profile of the MDR-0 group has a clinically substantial inferior profile with regard to three out of four most important criteria, cure rate, survival rate, TTD and probably also with regard to TTR, which is the last of the most important criterion. Furthermore, patients in the MDR-0 group have an inferior profile for

the medium-importance criterion mucositis/stomatitis and probably inferior to non-inferior on the low-importance criterion fatigue. Patients in MDR-1 have an inferior profile for myocardial ischemia and probably also an inferior profile for infections and arthralgia/myalgia.

The combination of genetic polymorphisms for patients in the MDR-1 group cause clinically significant and relevant differences for several clinical aspects, and the overall benefit-risk profile is positive. Based on these findings, it is reasonable to claim that genetic profiling is advisable in patients with colorectal cancer, to enable individualised treatment and follow-up. Genetic profiling will ensure that patients belonging to the MDR-1 group are identified and intensively monitored with regard to myocardial ischemia and infections that might impose a threat to the patient's health and can hereby be avoided. Patients in the MDR-0 group, on the other hand, can be carefully monitored for severe cases of mucositis/stomatitis, fatigue, etc. However, larger studies involving further patients are needed to verify the findings of this study.

6.3.5 Discussion and conclusion of colorectal case

The literature presents contradicting results with respect to the association of single polymorphisms and to TTR and TTD. Theoretically, however, the 5-FU metabolic phenotype is better explained by multi-gene and pathway-oriented analysis rather than by single gene analysis[117;118;120;121]. Data from the study by Afzal and co-workers was analysed using the same MDR classifications, and it was further preceded by dichotomously classifying toxicity into none-to-moderate (grade 0–2) and severe (grade 3–4) toxicity according to the MDR classification.

While supported by statistical analysis, the method itself is of a qualitative nature in order to properly allow for uncertainties and differences in opinion. This qualitative feature also serves to focus the assessment on clinical and toxicological issues and it allows comparison of benefits and risks that have largely different probability and/or significance in a justifiable detail.

Statistically significant results are essential, as discussed earlier in this thesis, but there is often insufficient attention to the clinical significance and relevance of data. It has previously been proposed that clinical significance is defined and justified based on the proportion of patients experiencing a specified treatment effect, and it has been demonstrated how clinical significance can be quantified and discussed in a qualitative manner.

The difference distribution captures even the smallest change or difference, and one could discuss whether patients surviving a day or a week more should be accounted for. This brings about very interesting discussions, which is probably the strongest advantage in using a structured framework for the benefit-risk assessment of different options. If desired, a cut-off point could be defined enabling, e.g. only patients that survive more than 1 month to be accounted for. The method can be tailored to the specific scenario, situation and decision context.

It is not the purpose of this study to discuss the strengths and shortcomings of the MDR classification, or the data in general. For such discussions, the reader is referred to the work by Afzal et al.[117;118;120;121]. It has been possible to demonstrate that the specific combination of polymorphism found by Afzal et al. causes a clinically significant and relevant increase in TTD and probably also in TTR. The differences between the groups are quantified and evaluated in a qualitative manner. Furthermore, tendencies in toxicity with regard to, e.g. an increased probability of severe myocardial ischemia in patients in the MDR-1 group have been demonstrated.

6.4 The schizophrenia case

All data included in this assessment is provided by Dr Gesche Jürgens and Dr Stig Ejdrup Andersen, Bispebjerg Hospital, Department of Clinical Pharmacology. The results are to be part of the health technology assessment prepared by Doctors Jürgens and Andersen for the National Board of Health, Denmark. The conclusions of this assessment are those of the author and are not to be attributed to the Department of Clinical Pharmacology and Department of Psychiatry, Bispebjerg Hospital.

6.4.1 Material

The cohort consists of 311 patients diagnosed with schizophrenia. All patients are tested with regard to their genotype of CYP2D6 and CYP2C19, and are divided into fast metabolisers, normal metabolisers and slow metabolisers. CYP2D6 and CYP2C19 play an important role in the metabolism of antipsychotics in humans. Patients are followed up to 400 days. Table 16 shows the clinical data from the study population.

They are divided into three treatment groups: 1) open CYP test, 2) blinded CYP test, but intensified clinical monitoring, and 3) blinded CYP test, control group. Only patients in the control group (Z) were revealed, and the rest were blinded with regard to the disposition into the two other treatment groups (groups X and Y).

Table 15: Clinical data on study cohort.

	N = 311
Age (median years, range)	41 (19-73)
Sex	
Male	172
Female	139

6.4.2 The method

1 Decision context: The aim of this assessment is to clarify whether individual dosing supported by genotyping improves the medical treatment with antipsychotics. A head-to-head comparison on the basis of the degree of severity of delusions, hallucinations, and main adverse events will be performed. Treatment groups X and Y will be scored against the control group, Z. If both groups perform better than the control group, a head-to-head comparison between them will be undertaken.

2 Decision profile: Based on the decision context, the following criteria are chosen to clarify the decision context: severity of delusions and hallucinations, and severity of main adverse event. A total of 74 criteria were found to be necessary for this assessment (see Figure 33). The selected criteria are justified to represent the most relevant aspects in the evaluation of differences between options in the assessment of schizophrenia.

3 Weighting: In this assessment, weighting is performed from a clinical point of view. The weights of adverse drug reactions were assigned by Dr Merete Nordentoft, Bispebjerg Hospital, Department of Psychiatry. The weights for delusions and hallucinations are assigned by Dr Gesche Jürgens.

Differences in high-importance criteria (weight 3) will have major clinical implications. Medium-importance criteria, which often are difficult to treat, may have clinical implications. Low-importance criteria (weight 1) are easy to handle and will rarely lead to clinical implications.

4 Scoring: All criteria are events and are scored using the confidence interval scoring method. The confidence level is set at 2/3. Adverse drug reactions were recorded and graded according to a severity scale of 0-3, where 0 = none, 1 = slight, 2 = moderate, and 3 = severe. Delusions and hallucinations were recorded and graded according to a severity

scale of 0-5, where 0 = none, 1 = questionable, 2 = mild, 3 = moderate, 4 = marked, and 5 = severe.

However, not all criteria were recorded for all patients, which induced problems in the creation of a single *scoring table* containing all criteria. Instead, individual scoring tables are created for each criterion. Based on baseline and follow-up values, the patients are divided into three groups: (i) responders, (ii) non-responders and (iii) adverse responders. These two values are used to create a difference value in the severity of disease. If a patient, e.g. at the baseline has a severity score of 3 for a criterion and a severity score of 2 at follow-up, the difference is -1. Using a scale, where a difference score between -1 to +1 is classified as a non-responder, the patient is classified as a non-responder. This crude definition is justified based on the type of data, patients, and questionnaire involved and used in this study.

Once all patients are divided into the three groups, a score is calculated for the responders and adverse responders, and two tornado plots are consequently created, one for the responder group and one for the adverse-responder group.

5 Evaluation of uncertainty: The evaluation of uncertainty and evidence is done qualitatively, where the objective scores of borderline criteria are changed to interval scores.

6 Weighted scores: The weights and scores are multiplied to gain weighted scores, which are visualised in step 7. Due to the overwhelming number of criteria, the standard table containing criteria, weights, scores and weighted scores is omitted in this case. The information about criteria, weights and scores is self-evident in the tornado diagrams (see e.g. Figure 33).

7 Visualisation:

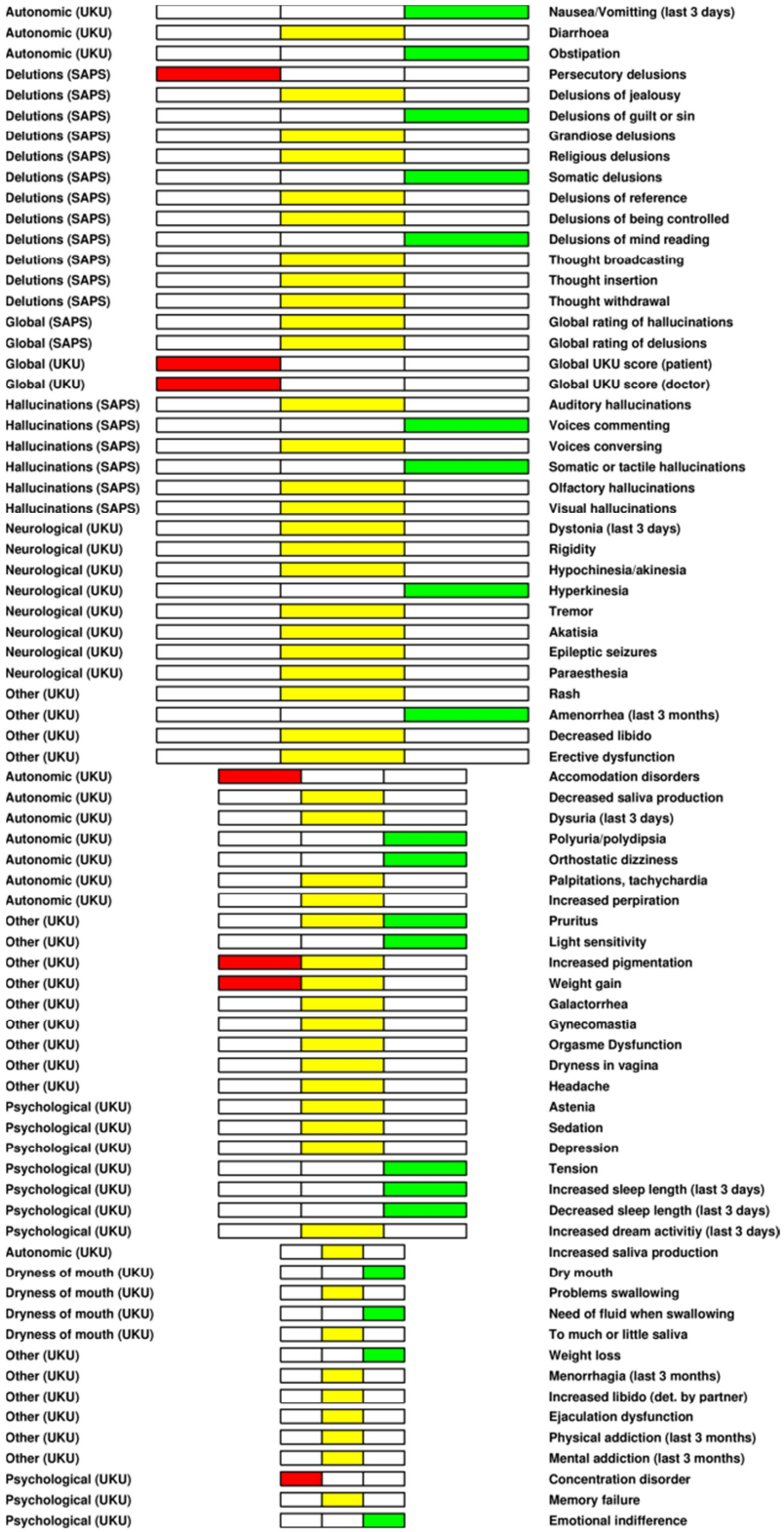


Figure 33: Tornado diagram showing the difference between treatment group X and the control group Z for responders. Criteria are listed according to their weights. To the right of the diagram, listed is the group to which the individual criterion belongs. Treatment group X is only inferior on few criteria in each weight class, while superior for multiple criteria in each weight class. The benefit-risk balance is positive.

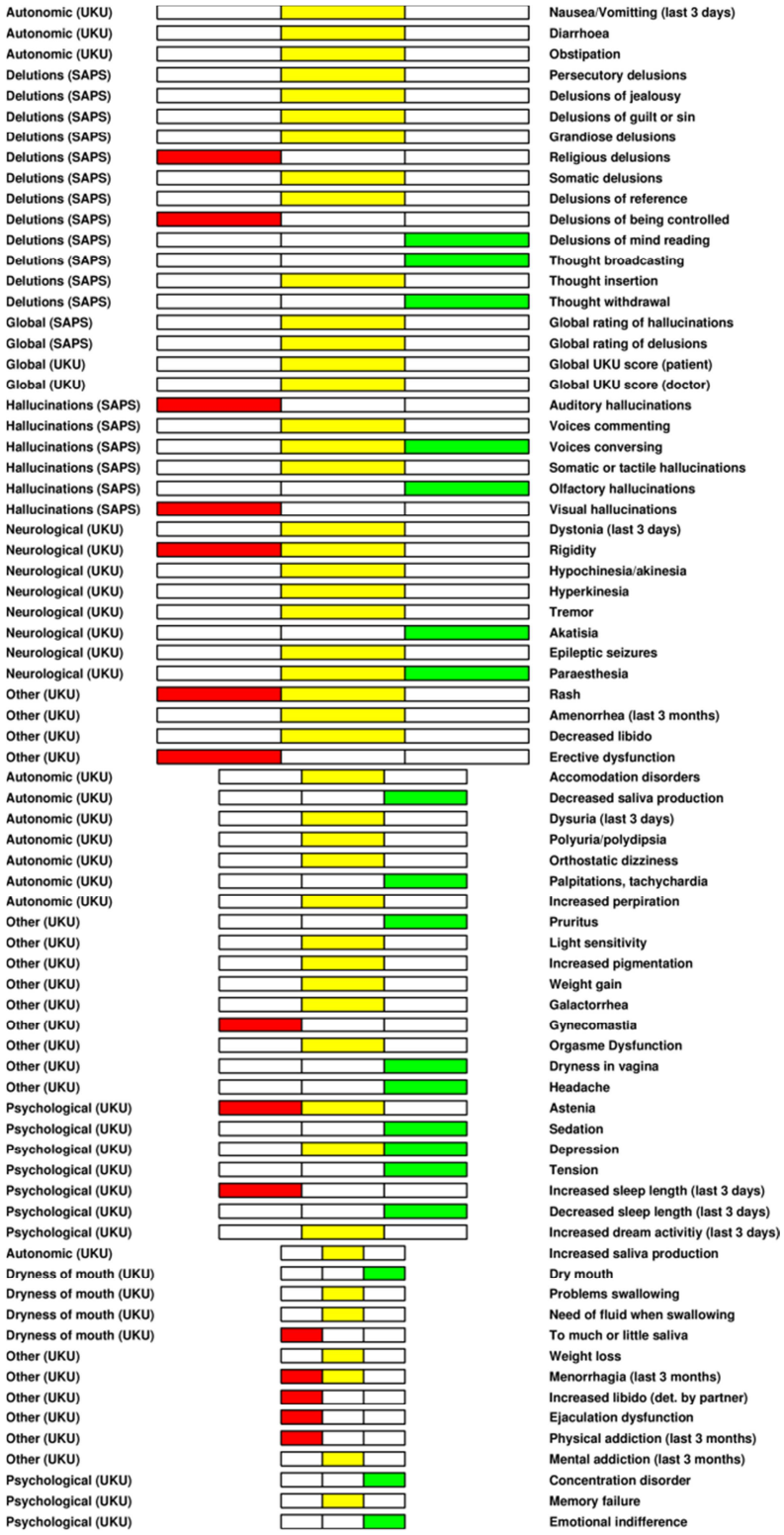


Figure 34: Tornado diagram showing the difference between treatment group X and the control group Z for adverse responders. Treatment group X is inferior on several criteria in each weight class, but also superior for several criteria in each weight class. The benefit-risk balance is deemed equal between treatment groups X and Z.

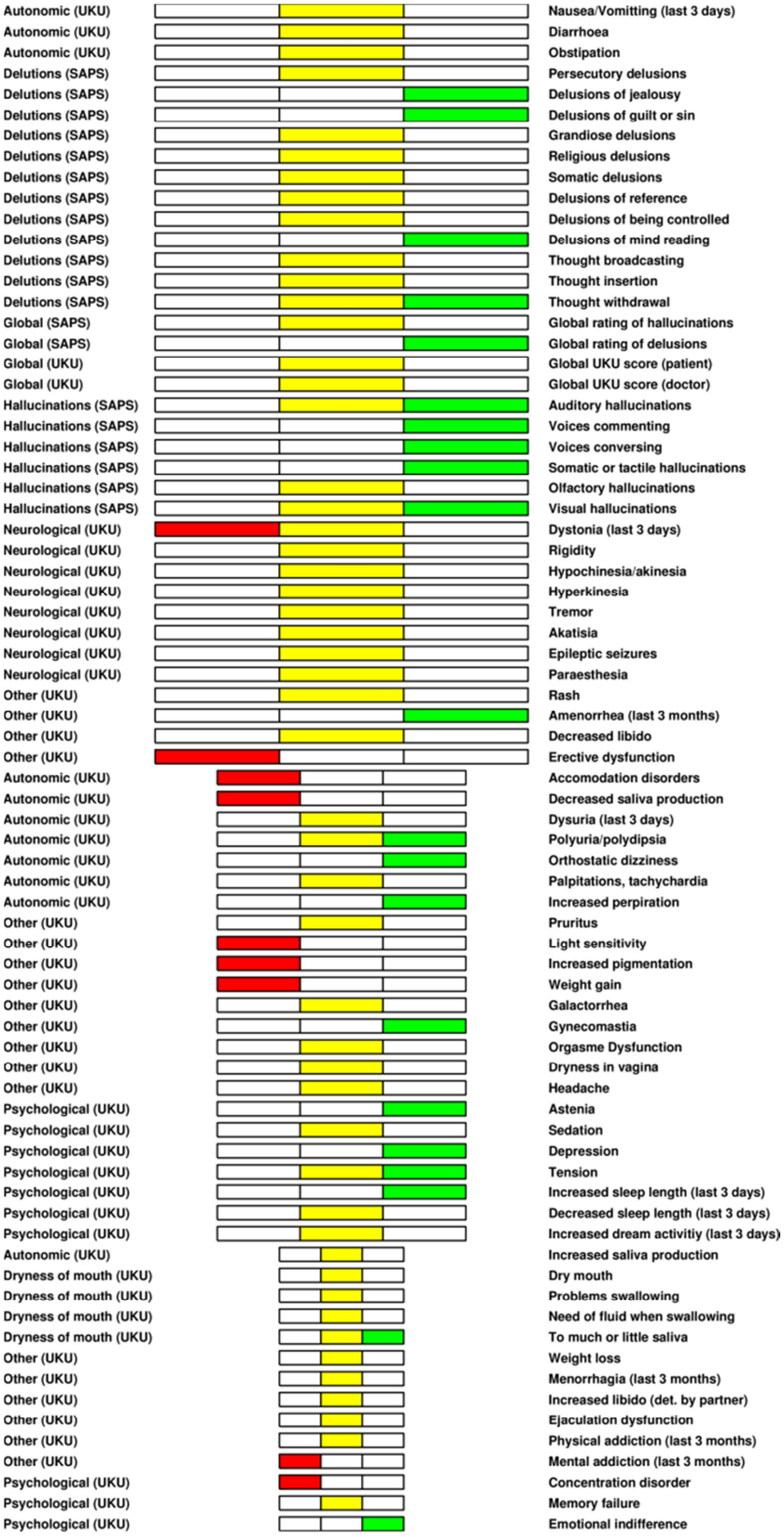


Figure 35: Tornado diagram showing the difference between treatment group Y and the control group Z for responders. Treatment group Y is superior on several criteria in each weight class, and only inferior for few criteria in each weight class. The benefit-risk balance is deemed positive in favour of treatment group Y.

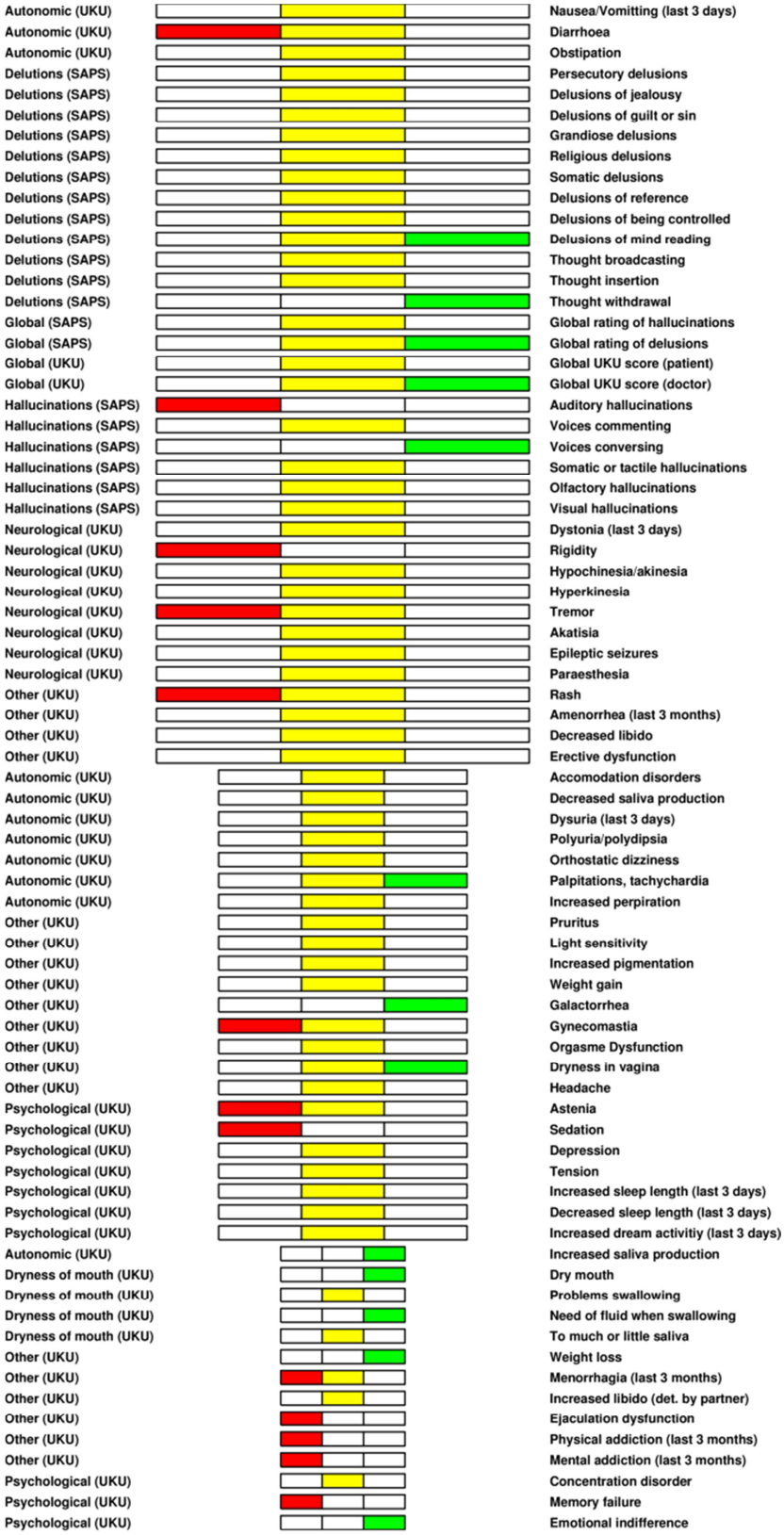


Figure 36: Tornado diagram showing the difference between treatment group Y and the control group Z for adverse responders. Treatment group Y is inferior on several criteria in each weight class, while also superior for several criteria in each weight class. The benefit-risk balance is deemed equal between the control group and treatment group Y.

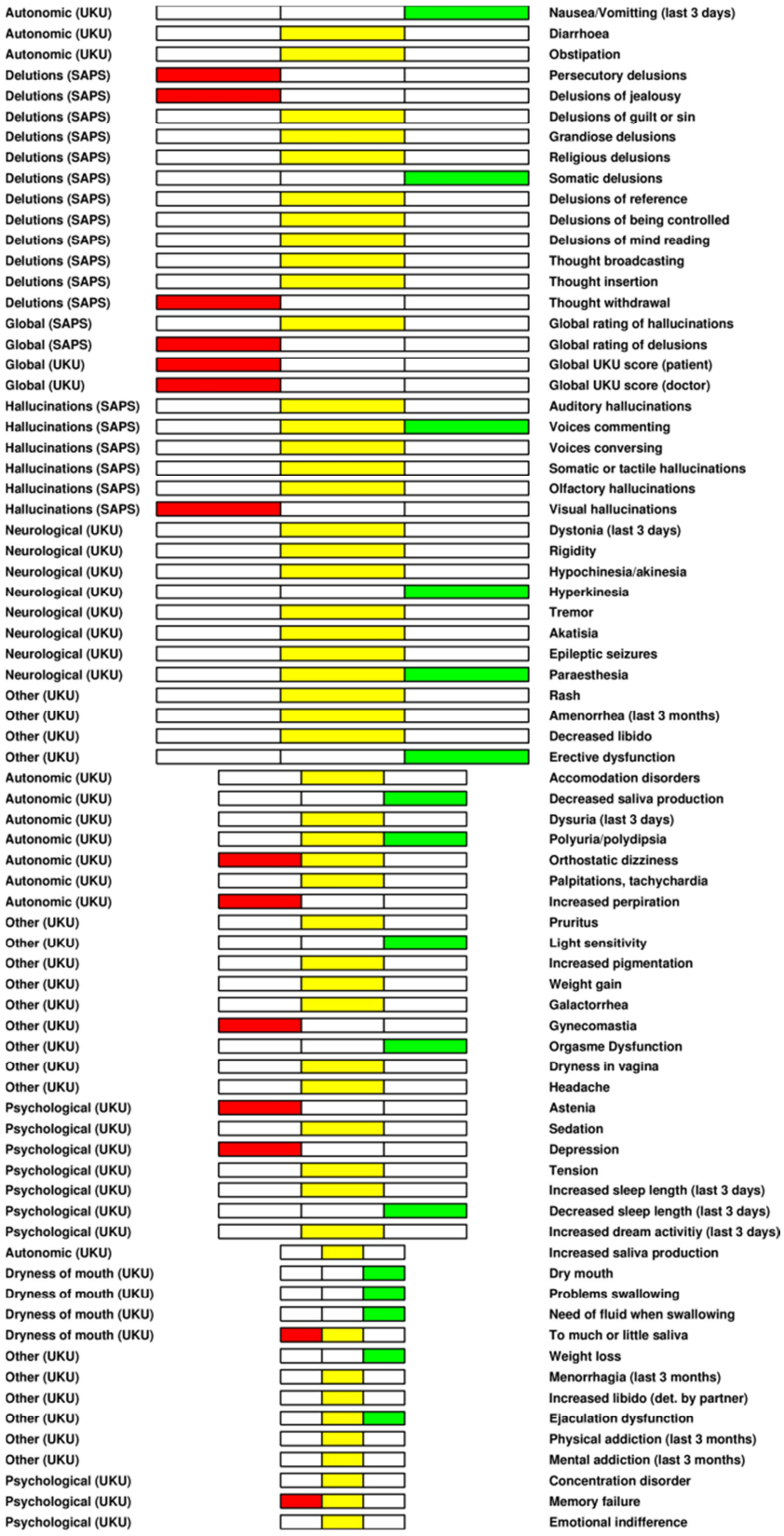


Figure 37: Tornado diagram showing the difference between treatment groups X and Y for responders. Treatment group Y is the comparator, and the scores shown are those of treatment group X. Treatment group X is inferior on several high-importance criteria, while also superior for several criteria in each weight class.

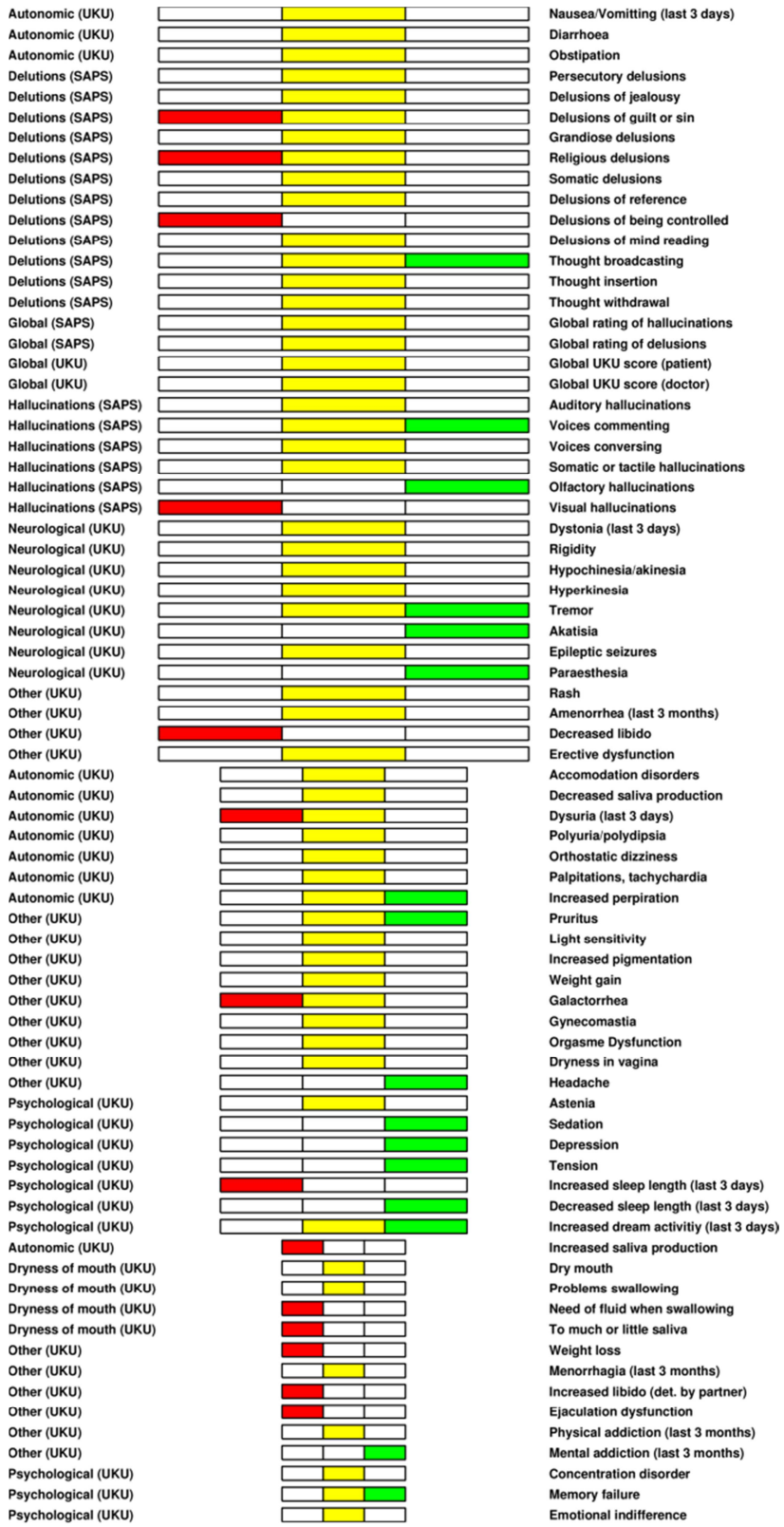


Figure 38: Tornado diagram showing the difference between treatment group X and the treatment group Y for adverse responders. Treatment group Y acts as the comparator, hence the scores for treatment group X are visualised. Treatment group X is inferior on several criteria in each weight class, while also superior for several criteria in each weight class. However, for the low-importance criteria,

treatment group X is inferior for multiple criteria, while only superior for one criterion. This inferior profile is, however, outweighed by the superior profile for medium-importance criteria. The benefit-risk balance is deemed equal between treatment group X and group Y.

8 Overall conclusions: The assessment shows that treatment groups X and Y perform better relative to the control group with regard to the responders, and equally with regard to adverse responders. However, it is interesting to notice that in the comparison of treatment group X relative to the control group Z although group X is superior for many criteria, group X fares worse than the control group with regard to the global score for adverse events (Global UKU score patient/doctor). This could be explained by the fact that, e.g. a male patient is doing very well with regard to many criteria, but is experiencing erectile dysfunction, which might be devastating for him.

The next step is to compare treatment groups X and Y, where one group acts as the comparator. This is done as seen in Figure 37 and 38, where treatment group Y acts as the comparator. For the responders, treatment group X is inferior for the high-importance criteria global ratings of delusions, and for global ratings of adverse drug reactions (Global UKU) both by patient and doctor.

In my view, even though data is blinded, it is not difficult to conclude that the clinical implications of these results are that treatment group Y may be favoured relative to treatment group X and control group Z. Apart from this, both treatment groups X and Y fare equally with regard to responders and adverse responders, and the overall benefit-risk balance of treatment group X is non-inferior, although a small but positive tendency towards treatment group Y is seen. The stakeholders in this assessment should, in my opinion, only have assigned the weight “high” (3) to the global criteria, such as global rating of delusions, hallucinations and adverse events (UKU), to clearly indicate the importance of these criteria. Further studies are needed to confirm these findings.

6.5 The Ketek® case

All information and data used for this assessment is publicly available on EMA's website[122]. The aim and goal of this assessment is to demonstrate how the conclusions made in the EPAR can be communicated and visualised by the use of the method described in Chapter 5.

6.5.1 Background and Material

The active substance in Ketek is telithromycin. Telithromycin is a semi-synthetic derivative of erythromycin A, belonging to ketolide antibiotics, a class related to macrolides. Ketek is indicated in the treatment of community-acquired Respiratory Tract Infections (RTI), Community-acquired pneumonia, mild or moderate (CAP), acute exacerbation of chronic bronchitis (AECB), acute sinusitis (ABS), and tonsillitis/pharyngitis (TP). Beta-lactam agents and macrolides are commonly used for the treatment of community-acquired RTI, although resistance against *S. pneumoniae* has reached significant levels in several European countries.

Efficacy was evaluated in 16 phase III studies (12 double-blind comparative trials and 4 open-label non-comparative trials) with endpoints cure or failure of cure. For the indication AECB, one study - a phase IV trial - was presented with the primary endpoint being percentages of patients harbouring a penicillin- or erythromycin-resistant S. pneumoniae (PERSp) at the "test of cure (TOC) visit" amongst patients with Streptococcus pneumoniae infections at inclusion.

Efficacy data from phase III trials, including open-label and phase IV trials, are pooled in this assessment, analysing all telithromycin treated versus all comparators treated, for each indication. This will be done to the extent possible based on the information available in the EPAR, which does not include all data from phase IV studies. Safety data are already pooled in the EPAR, although pooled phase III double-blinded, pooled open-label and pooled phase IV data are given separately.

6.5.2. The method

1 Decision context: The aim of this assessment is to judge whether the benefit-risk profile of Ketek is satisfactory for a marketing authorisation in the following indications, and whether any of these indications should be approved with restrictions:

In patients of 18 years and older:

- Community-acquired pneumonia, mild or moderate (CAP)
- Acute exacerbation of chronic bronchitis (AECB)
- Acute sinusitis (ABS)

In patients of 12 years and older:

- Tonsillitis/pharyngitis (TP) caused by Group A *beta streptococci*, as an alternative when beta-lactam antibiotics are not appropriate

The assessment will be made with the perspective of the regulators. Several alternative treatments are registered for the four indications:

- CAP: Amoxicillin, Clarithromycin, Trovafloxacin
- AECB: Amoxicillin-clavulanic acid, Cefuroxime, Clarithromycin, Azithromycin
- ABS: Amoxicillin-clavulanic acid and Cefuroxime
- TP: Penicillin, clarithromycin

All alternatives, except Azithromycin, have been used as comparators in the different phase III trials for the relevant indications.

2 Decision profile: Based on the decision context, the following criteria are chosen: Cure, PERSp at TOC (for the indication AECB), treatment-emergent adverse events (TEAEs), serious adverse events (SAEs), hepatic adverse events (AEs), cardiac AEs, visual AEs, and syncope.

3 Weighting: In this study, weighting is performed from a regulatory point of view. Cure is the primary endpoint, since patients, clinicians and regulators are interested in overall cure. Furthermore, serious adverse events (SAE) and hepatic AEs are weighted as high, since several other antibiotics for the same indications exist. Therefore, any tendency showing an unfavourable number of SAEs and/or hepatic events for Ketek will be highly concerning. These criteria are therefore of high importance, meaning that a difference between Ketek and a comparator will have major consequences.

The criteria syncope, cardiac events and PERSp are all considered medium important, meaning that a difference between two options will probably have clinical and regulatory implications, e.g. if the performance for the high-importance criteria is equal between the two options. The medium-importance criteria are often difficult to treat and can have consequences for the patients, occasionally very serious to fatal.

The low-importance criteria, visual adverse events and TEAE can be very severe, but they are rarely life threatening and can easily be treated.

4 Scoring: All data describing the stated criteria are events and are scored using the confidence interval scoring method. The confidence level is set at 2/3.

5 Evaluation of uncertainty: The evaluation of uncertainty and evidence is done qualitatively, where the objective scores of borderline criteria are changed to interval scores.

6 Weighted scores:

Criterion	Weight	Score				Weighted Score			
		CAP	ABS	AECB	TP	CAP	ABS	AECB	TP
Cure	3	+1	1	0	0	+3	3	0	0
PERSp	2	-	-	1 0	-	-	-	2 0	-
Syncope	2	**0**-1	**0**-1	-1	0	**0**-2	**0**-2	-2	0
Visual Events	1	**-1**-0	-1	-1	-1	**-1**-0	-1	-1	-1
Cardiac Events	2	0-**1**	**0**-1	**0**-1	0	0-**2**	**0**-2	**0**-1	0
SAE	3	1	**-1**-0	-1	0	3	**-3**-0	-3	0
Hepatic Events	3	1	-1	0	+1	3	-3	0	+3
TEAE	1	1	-1	-1	-1	1	-1	-1	-1

For interval scores, the bold numbers are the objective scores based on the score table.

7 Visualisation:

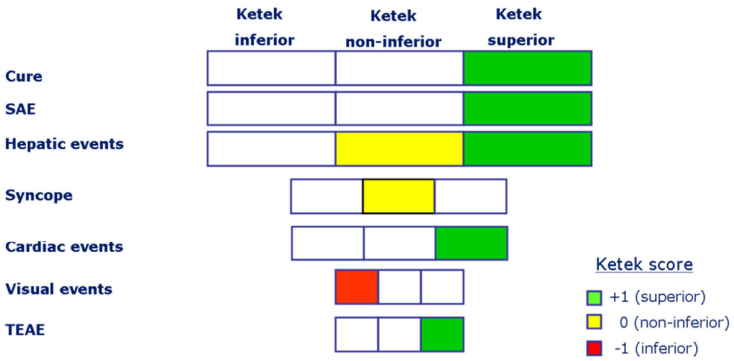


Figure 39: The main results for the indication CAP. Ketek is only inferior on the criterion “visual events”.

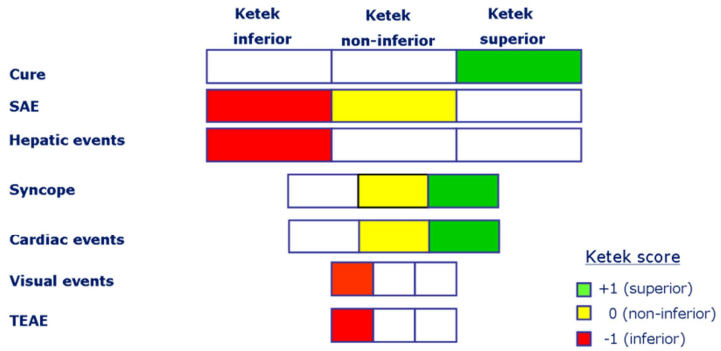


Figure 40: The main results for the indication ABS. Ketek is inferior on two out of three high-importance criteria and all low-importance criteria. Ketek is only superior on the high-importance criterion “cure”.

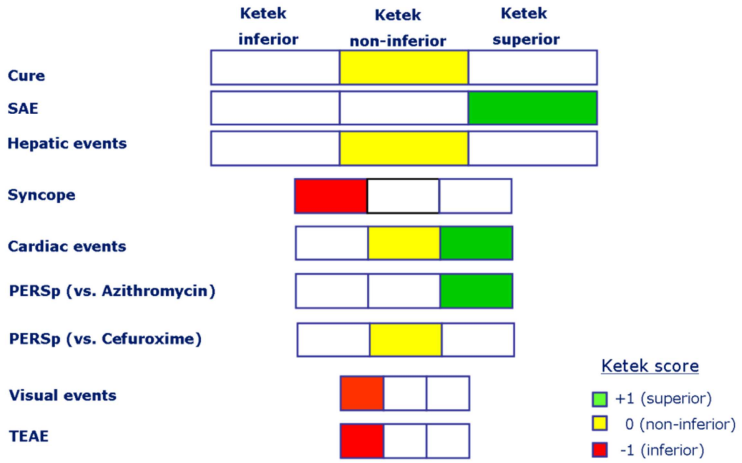


Figure 41: The main results for the indication AECB, where Ketek is inferior on the criteria syncope, visual events and TEAE. Ketek is only superior for one of the high-importance criteria, SAE, and one of the medium-importance criteria, PERSp (vs. Cefuroxime).

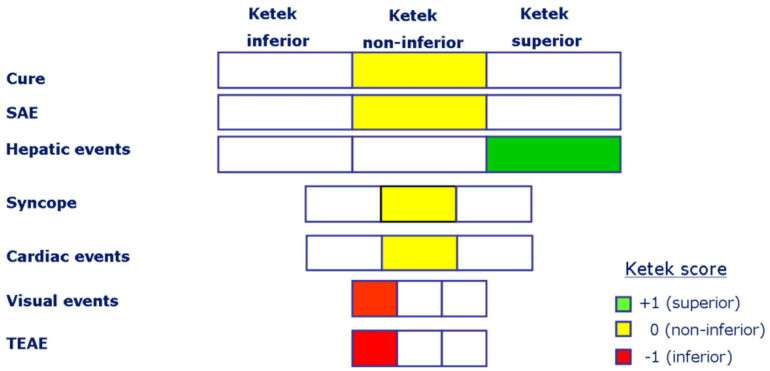


Figure 42: The main results for the indication TP, where Ketek is only superior on the high-importance criterion “hepatic events”. For the remaining criteria, Ketek is either non-inferior or inferior.

8 Overall conclusions: For the indication community-acquired pneumonia, mild or moderate (CAP), the high-importance criteria and cure and SAE were found to be in favour of Ketek, while the high-importance risk criterion hepatic AEs is non-inferior to superior. Ketek is also superior to comparators on cardiac AEs and TEAEs, while non-inferior for the criterion syncope. The only criterion that is in favour of comparators is visual AEs, which is judged to be of lower importance in this context. Overall, the benefit-risk balance of Ketek is considered positive.

For the indication acute sinusitis (ABS), the high-importance criterion cure was found to be in favour of Ketek. However, for high-importance criterion hepatic AEs, Ketek was found to be inferior, while inferior to non-inferior for SAE. Furthermore, the risk criteria visual AEs and TEAEs were in favour of the comparator, while the analyses of the criteria syncope and cardiac AEs showed non-inferiority to superiority. Overall, the benefit-risk balance of Ketek is not considered to be positive.

Considering the indication acute exacerbation of chronic bronchitis (AECB), Ketek was found to be non-inferior for the high-importance criterion cure. Ketek is inferior on the criteria syncope, visual events and TEAE. Ketek is only superior for one of the high-importance criteria, SAE, and one of the medium-importance criteria, PERSp (vs. Cefuroxime). For this indication, there is also data from a phase IV trial on penicillin- or erythromycin-resistant S. pneumoniae (PERSp) at TOC (test of cure). For this criterion, Ketek is superior when compared to Azithromycin, and non-inferior when compared to Cefuroxime.

With regard to the indication tonsillitis/pharyngitis (TP), the high-importance criterion hepatic AEs is the only criterion that shows a positive tendency towards Ketek, while both low-importance criteria, visual AEs and TEAE, are in favour of the comparators. The benefit-risk profile for Ketek is considered to be compatible with comparators.

In conclusion, Ketek has a non-inferior to inferior profile for most criteria relative to other comparators, but a superior profile with regard to cure and resistance. Therefore, the use of Ketek is advisable in patients where there is a failure of other treatments either due to a lack of efficacy or risk of bacterial resistance.

6.6 Discussion and conclusion of chapter 6

The proposed methodology was developed as an iterative process by testing the method in four different workshops. This approach not only enabled the development of a tailored method to the needs expressed by project teams, but also enabled the development of a tailored method for the benefit-risk assessment of clinical data from drugs in development. The method was presented at the Danish Medicines Agency and was seen as a possible good supplement to the current assessments conducted by pharmaceutical companies. However, it was also expressed that the method needed validation first. I believe that the concept of validation is misused in the field of benefit-risk assessment. This is mainly due to the fact that the general methods presented and discussed in this thesis, including the current methodology, have subjective input, which is impossible to validate. The concept of validation belongs to, e.g. statistical methods with no subjective input. Therefore, this term must be avoided in the field of benefit-risk assessment. The only “validation” a benefit-risk assessment can get is the acceptance it can get from regulatory agencies through its use by pharmaceutical companies.

The method was tested on several occasions, but for reasons of confidentiality, results from Novo Nordisk A/S are not presented. On several occasions, the method was used and tested on external data. A presentation of the results gained from the colorectal, schizophrenia and telithromycin (Ketek®) case have been presented.

7 Final discussion and conclusions

There is no generally accepted framework for the benefit-risk assessment of medicines and treatments. The development costs of a new medicine often exceed 1 billion USD and often take more than 10 years to develop. Therefore, there is an obvious need for a structured and transparent benefit-risk assessment methodology that is specifically tailored for the drug development and approval processes. In this context, structured implies that the assessment should follow a standardised set of steps, each with a clearly defined purpose. Transparent means that all assumptions made in the processes of firstly choosing benefit-risk criteria and thereafter scaling and weighting these criteria must be clearly expressed all the way to the final decision. This is important both to avoid that unreported biases and feedback distort the assessment and to make it possible for the industrial partner and the regulatory agency to compare the results of their evaluation on a point-by-point basis. Ideally, the method should also be dynamic in the sense that the assessment can be updated regularly to account for new information about the functioning or side effects of the drug.

The method should be based on data from experimental and clinical studies as well as from other sources of information. While supported by statistical analysis, the method itself should be of a qualitative nature in order to properly allow for uncertainties and differences in opinion. This qualitative feature also serves to focus the assessment on clinical and toxicological issues and it allows comparison of benefits and risks that have largely different probability and/or significance in a justifiable detail.

The benefit-risk assessment methodology presented in this thesis has been specifically tailored in view of overcoming these difficulties. It is inspired by Multi-Criteria Decision Analysis (MCDA). However, the method offers a simpler approach to integrate information of many different types, and it provides a clear diagrammatical format for presenting the essential decision criteria, their importance, and the degree to which a given drug satisfies the different criteria.

Different stakeholders, e.g. industry and regulatory agencies, are likely to have different views on weighting and scoring of the various criteria and they may also disagree about the choice of decision criteria. It is therefore important that this information is fully maintained and clearly presented in the final decision process. On the other hand, it is also clearly important that the underlying statistical analysis is available in a clearly understandable form.

A comprehensive approach to benefit-risk assessment for a new drug, focusing on transparency, clinical significance and relevance of data, the visualisation and communication of results, has been proposed and described. The method has been validated by several different examples clearly demonstrating its general applicability. All examples include a qualitative and a quantitative part, and they combine these aspects in a simple and transparent fashion.

The method can handle a variety of different types of data encountered in clinical trials and can be used in single trials as well as in multiple trials. In principle, it can also be developed into a dynamic assessment that follows a drug from its first conception to the end of its life. A comparison with other comprehensive methods, e.g. MCDA, BRAT or CIRS, would have been desirable, but unfortunately due to limited time, this could not be accomplished. However, the method should be evaluated in comparison with other methods based on several critical issues, e.g. logical soundness, comprehensiveness and practicality, to fully reveal its strengths and limitations. A pilot study at a regulatory authority and a pharmaceutical company would resolve this issue.

None of the existing comprehensive approaches focus on patient perspective, economical aspects, etc. A desire to involve these aspects in the benefit-risk assessment is nearby and appealing. A new medicine or treatment may, e.g. have economical and ethical consequences that need to be addressed as early as in the approval phase by regulatory agencies.

Currently, patient perspective is not systematically included in drug development or approval phase. A structured approach taking the patient perspective into account is lacking, both in the industry and the regulatory environment. Drugs are developed for patients, and it

is natural to conclude that patients should be in the centre of any decision made regarding the benefit-risk assessment of a medicine/treatment. However, this is far from the current practice, and a shift in the paradigm is much needed.

Appendix A - Questions used in the interviews with the DMA

Use of methods

1. Do you know of a validated and/or structured process, even just a checklist for assessing benefit-risk assessment except for D80 AR?
2. Do you experience a need for a structured process in your work/assessment?

How can benefit-risk assessment be improved? (The following questions are inspired by EMA's work on benefit-risk assessment)

1. How can the pharmaceutical industry improve the quality of benefit-risk assessments?
2. Does the agency encourage the use of structured benefit-risk assessment?
3. How can the pharmaceutical industry improve transparency and consistency of benefit-risk assessment?
4. How can the pharmaceutical industry conduct more auditable and robust benefit-risk assessments?
5. How can the agency and the industry harmonise benefit-risk assessment?

The benefit-risk assessment method:

1. Which elements of the method can improve current benefit-risk assessment?
2. How can the method be improved to better fulfil the agencies' expectations to benefit-risk assessment?
3. Would the method result in more robust assessments?
4. Can the method be used to harmonise benefit-risk assessment between the industry and the agency? If NO, why? If YES, how?
5. Do you encourage Novo Nordisk A/S to use the method or elements of it in meetings with the agency and/or in a marketing application?

Reference list

1. The Council for International Organizations of Medical Sciences (CIOMS). Working CIOMS Group IV - Benefit-Risk Balance for Marketed Drugs: Evaluating Safety Signals [online]. Available from URL: http://www.cioms.ch/publications/g4-benefit-risk.pdf . 1998.

2. Couzin J. Drug safety. Withdrawal of Vioxx casts a shadow over COX-2 inhibitors. *Science*. 2004 Oct 15;306(5695):384-5.

3. Misbin RI. Lessons from the Avandia controversy: a new paradigm for the development of drugs to treat type 2 diabetes. *Diabetes Care*. 2007 Dec;30(12):3141-4.

4. Honig P, Lalonde R. The economics of drug development: a grim reality and a role for clinical pharmacology. *Clin Pharmacol Ther*. 2010 Mar;87(3):247-51.

5. Mussen F, Salek S, Walker S. 978-0-470-06085-8 (H/B) Benefit-Risk Appraisal of Medicines - A systematic approach to decision-making. Wiley-Blackwell; 2009.

6. US Food and Drug Administration (FDA). The Future of Drug Safety - Promoting and Protecting the Health of the Public [online]. Available from URL: http://www.fda.gov/downloads/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/UCM171627.pdf . 2007.

7. Identifying CDER's Science and Research Needs Report - The CDER Science Prioritization and Review Committee (SPaRC). [online] Available from URL: http://www.fda.gov/downloads/drugs/scienceresearch/ucm264594.pdf. 2011.

8. FDA: PDUFA V. Available online: http://www.fda.gov/downloads/ForIndustry/UserFees/PrescriptionDrugUserFee/UCM270412.pdf. 2011.
Ref Type: Online Source

9. Committee for Medicinal Products for Human Use (CHMP). Report of the CHMP working group on benefit-risk assessment models and methods. EMEA/CHMP/15404/2007 [online]. Available from URL: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2010/01/WC500069634.pdf. CHMP; 2008.

10. Benefit-risk methodology project - work package 2 report: Applicability of current tools and processes for regulatory benefit-risk assessment. EMA/549682/2010 [online]. Available from URL: http://www.ema.europa.eu/docs/en_GB/document_library/Report/2010/10/WC500097750.pdf . European Medicines Agency (EMA); 2010.

11. Holden WL. Benefit-risk analysis: a brief review and proposed quantitative approaches. *Drug Saf*. 2003;26(12):853-62.

12 . Felli JC, Noel RA, Cavazzoni PA. A multiattribute model for evaluating the benefit-risk profiles of treatment alternatives. *Med Decis Making*. 2009 Jan;29(1):104-15.

13 . Goetghebeur MM, Wagner M, Khoury H, Levitt RJ, Erickson LJ, Rindress D. Evidence and Value: Impact on DEcisionMaking - the EVIDEM framework and potential applications. *BMC Health Serv Res*. 2008;8:270.

14 . Guo JJ, Pandey S, Doyle J, Bian B, Lis Y, Raisch DW. A review of quantitative risk-benefit methodologies for assessing drug safety and efficacy - report of the ISPOR risk-benefit management working group. *Value Health*. 2010 Aug;13(5):657-66.

15 . Levitan BS, Andrews EB, Gilsenan A, Ferguson J, Noel RA, Coplan PM, et al. Application of the BRAT Framework to Case Studies: Observations and Insights. *Clin Pharmacol Ther*. 2011 Feb;89(2):217-24.

16 . Mussen F, Salek S, Walker S. A quantitative approach to benefit-risk assessment of medicines - part 1: the development of a new model using multi-criteria decision analysis. *Pharmacoepidemiol Drug Saf*. 2007 Jul;16 Suppl 1:S2-S15.

17 . Belton V, Stewart TJ. Multiple Criteria Decision Analysis - An Integrated Approach. Kluwer Academic Publishers; 2002.

18 . Keeney RL. Value Focused Thinking - A Path to Creative Decisionmaking. Harvard University Press; 1992.

19 . Dodgson J, Spackman M, Pearman A, Phillips L. DTLR multi-criteria analysis manual. Available online from URL: http://iatools.jrc.ec.europa.eu/public/IQTool/MCA/DTLR_MCA_manual.pdf. 1998.

20 . Keeney RL, Raiffa H. Decisions with Multiple Objectives - Preferences and Value Tradeoffs. 1976.

21 . Løken E. Use of multicriteria decision analysis methods for energy planning problems. *Renewable and Sustainable Energy Reviews*. 2007;11:1584-95.

22 . Dyk Ev, Smith DG. R&D Portfolio Selection by Using Qualitative Pairwise Comparisons. *OMEGA Int J of Mgmt Sci*. 1990;18(6):583-94.

23 . Walker S., McAuslane N, Liberti L. Refining the Benefit-Risk Framework for the Assessment of Medicines: Valuing and weighting benefit and risk parameters; Washington, DC: 17-18 June, 2010. Report of the Workshop organised by the CMR International Institute for Regulatory Science. Available from URL: http://www.cmr.org. 2010.

24 . Walker S., Liberti L. Visualising benefit-risk of assessment of medicines: The key to developing a framework that informs stakeholder perspective and clarity of decision making. 2011.

25 . Boudes PF. How to improve complex drug development? A critical review of FDA advisory meetings. *Drug Information Journal*. 2007;41(5):673-83.

26 . MILLER GA. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol Rev*. 1956 Mar;63(2):81-97.

27 . EMA D80 AR: Overview and list of questions. Available from URL: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2009/10/WC500004800.pdf. 2010.

28 . CHMP D80 AR clinical aspects. EMA/577696/2010Rev10.10. [online] Available from URL: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2009/10/WC500004856.pdf. 2010.

29 . CHMP GUIDANCE DOCUMENT ON VOTING IN THE FRAMEWORK OF DISCUSSION AND ADOPTION OF CHMP OPINIONS. CPMP/3137/01 Rev 1.1 [online] Available from URL: http://www.emea.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2009/10/WC500004181.pdf. 2008.

30 . FDA Guidance for Industry: Development and Use of Risk Minimization Action Plans (RiskMAP Guidance). Available from URL: http://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM126830.pdf. 2005.

31 . FDA Guidance for Industry: Premarketing Risk Assessment (Premarketing Guidance). Available from URL: http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm126958.pdf. 2005.

32 . FDA Guidance for Industry: Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment (Pharmacovigilance Guidance). Available from URL:http://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM126834.pdf. 2005.

33 . FDA Good Reviewer Practice: Clinical Safety Review of an NDA or BLA. Available online from URL: http://www.fda.gov/downloads/AboutFDA/CentersOffices/CDER/ManualofPoliciesProcedures/UCM236905.pdf. 2010.

34 . Benefit Risk Methodology Project - work package 1: Description of the current practice of benefit-risk assessment for centralised procedure products in the EU regulatory network. EMA/213482/2010 [online]. Available from URL: http://www.ema.europa.eu/docs/en_GB/document_library/Report/2010/04/WC500089603.pdf. 2010.

35 . Callréus T. The Precautionary Principle and Pharmaceutical Risk Assessment. *Drug Saf*. 2005;28(6):465-71.

36 . Ricci PF, Cox LAj, MacDonald TR. Precautionary principles: a jurisdiction-free framework for decision-making under risk. *Human & Experimental Toxicology*. 2004;23:579-600.

37 . Tallarida RJ, Murray RB, Eiben C. A scale for assessing the severity of diseases and adverse drug reactions. Application to drug benefit and risk. *Clin Pharmacol Ther*. 1979 Apr;25(4):381-90.

38 . Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med*. 1988 Jun 30;318(26):1728-33.

39 . Holden WL, Juhaeri J, Dai W. Benefit-risk analysis: a proposal using quantitative methods. *Pharmacoepidemiol Drug Saf*. 2003 Oct;12(7):611-6.

40 . Holden WL, Juhaeri J, Dai W. Benefit-risk analysis: examples using quantitative methods. *Pharmacoepidemiol Drug Saf*. 2003 Dec;12(8):693-7.

41 . Lynd LD, O'Brien BJ. Advances in risk-benefit evaluation using probabilistic simulation methods: an application to the prophylaxis of deep vein thrombosis. *J Clin Epidemiol*. 2004 Aug;57(8):795-803.

42 . Lynd LD, Marra CA, Najafzadeh M, Sadatsafavi M. A quantitative evaluation of the regulatory assessment of the benefits and risks of rofecoxib relative to naproxen: an application of the incremental net-benefit framework. *Pharmacoepidemiol Drug Saf*. 2010 Nov;19(11):1172-80.

43 . Riegelman R, Schroth WS. Adjusting the number needed to treat: incorporating adjustments for the utility and timing of benefits and harms. *Med Decis Making*. 1993 Jul;13(3):247-52.

44 . Grieve AP. The number needed to treat: a useful clinical measure or a case of the Emperor's new clothes? *Pharmaceutical Statistics*. 2003;2:87-102.

45 . Kristiansen IS, Gyrd-Hansen D, Nexoe J, Nielsen JB. Number needed to treat: easily understood and intuitively meaningful? Theoretical considerations and a randomized trial. *J Clin Epidemiol*. 2002 Sep;55(9):888-92.

46 . Altman DG, Machin D, Bryant TN, Gardner MJ. Statistics with confidence. 2 ed. BMJ Books; 2000.

47 . Hildebrandt M, Vervolgyi E, Bender R. Calculation of NNTs in RCTs with time-to-event outcomes: a literature review. *BMC Med Res Methodol*. 2009;9:21.

48 . Hutton JL. Number needed to treat: properties and problems. *J R Statist Soc A*. 2000;163(3):403-19.

49 . Altman DG, Jenkins JD. Comments on the paper by Hutton, J.L. *J R Statist Soc A*. 2003;163(3):415-6.

50 . Garrison LP, Jr., Towse A, Bresnahan BW. Assessing a structured, quantitative health outcomes approach to drug risk-benefit analysis. *Health Aff (Millwood)*. 2007 May;26(3):684-95.

51 . Garrison LP. Regulatory benefit-risk assessment and comparative effectiveness research: strangers, bedfellows or strange bedfellows? *Pharmacoeconomics*. 2010 Oct 1;28(10):855-65.

52 . Weinstein MC, Torrance G, McGuire A. QALYs: the basics. *Value Health*. 2009 Mar;12 Suppl 1:S5-S9.

53 . Thompson JP, Noyes K, Dorsey ER, Schwid SR, Holloway RG. Quantitative risk-benefit analysis of natalizumab. *Neurology*. 2008 Jul 29;71(5):357-64.

54 . Murray CJ. Quantifying the burden of disease: the technical basis for disability-adjusted life years. *Bull World Health Organ*. 1994;72(3):429-45.

55 . Wolfson MC. Health-adjusted life expectancy. *Health Rep*. 1996;8(1):41-6.

56 . Gelber RD, Goldhirsch A, Cole BF, Wieand HS, Schroeder G, Krook JE. A quality-adjusted time without symptoms or toxicity (Q-TWiST) analysis of adjuvant radiation therapy and chemotherapy for resectable rectal cancer. *J Natl Cancer Inst*. 1996 Aug 7;88(15):1039-45.

57 . Gelber RD, Lenderking WR, Cotton DJ, Cole BF, Fischl MA, Goldhirsch A, et al. Quality-of-life evaluation in a clinical trial of zidovudine therapy in patients with mildly symptomatic HIV infection. The AIDS Clinical Trials Group. *Ann Intern Med*. 1992 Jun 15;116(12 Pt 1):961-6.

58 . Revicki DA, Feeny D, Hunt TL, Cole BF. Analyzing oncology clinical trial data using the Q-TWiST method: clinical importance and sources for health state preference data. *Qual Life Res*. 2006 Apr;15(3):411-23.

59 . Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. *Ann Intern Med*. 2009 Aug 4;151(3):203-5.

60 . Garber AM, Tunis SR. Does comparative-effectiveness research threaten personalized medicine? *N Engl J Med*. 2009 May 7;360(19):1925-7.

61 . Horn SD, Gassaway J. Practice-based evidence study design for comparative effectiveness research. *Med Care*. 2007 Oct;45(10 Suppl 2):S50-S57.

62 . Ryan M, Farrar S. Using conjoint analysis to elicit preferences for health care. *BMJ*. 2000 Jun 3;320(7248):1530-3.

63 . Ryan M. Using conjoint analysis to take account of patient preferences and go beyond health outcomes: an application to in vitro fertilisation. *Soc Sci Med*. 1999 Feb;48(4):535-46.

64 . Edwards R, Wiholm BE, Martinez C. Concepts in risk-benefit assessment. A simple merit analysis of a medicine? *Drug Saf*. 1996 Jul;15(1):1-7.

65 . Howard RA. Decision Analysis: Practice and Promise. *Management Science*. 1988;34(6):679-95.

66 . Guitouni A, Martel J. Tentative guidelines to help choosing an appropriate MCDA method. *European Journal of Operational Research*. 1998;109:501-21.

67 . Eriksen S, Keller LR. A multiattribute-utility-function approach to weighing the risks and benefits of pharmaceutical agents. *Med Decis Making*. 1993 Apr;13(2):118-25.

68 . Khan AA, Perlstein I, Krishna R. The use of clinical utility assessments in early clinical development. *AAPS J*. 2009 Mar;11(1):33-8.

69 . Poland B, Hodge FL, Khan A, Clemen RT, Wagner JA, Dykstra K, et al. The clinical utility index as a practical multiattribute approach to drug development decisions. *Clin Pharmacol Ther*. 2009 Jul;86(1):105-8.

70 . Ouellet D. Benefit-risk assessment: the use of clinical utility index. *Expert Opin Drug Saf*. 2010 Mar;9(2):289-300.

71 . Ouellet D, Werth J, Parekh N, Feltner D, McCarthy B, Lalonde RL. The use of a clinical utility index to compare insomnia compounds: a quantitative basis for benefit-risk assessment. *Clin Pharmacol Ther*. 2009 Mar;85(3):277-82.

72 . Renard D, Wu K, Wada R, Flesch G. Using desirability indices for decision making in drug development. Available online from URL: http://www.page-meeting.org/pdf_assets/4163-PAGE2009%20poster_Renard.pdf. 2009.
Ref Type: Online Source

73 . Chuang-Stein C, Mohberg NR, Sinkula MS. Three measures for simultaneously evaluating benefits and risks using categorical data from clinical trials. *Stat Med*. 1991 Sep;10(9):1349-59.

74 . Chuang-Stein C, Entsuah R, Pritchett Y. Measures for Conducting Comparative Benefit:Risk Assessment. *Drug Information Journal*. 2008;42:223-33.

75 . Walker S., Phillips L, Cone M. Benefit-Risk Assessment Model for Medicines: Developing a Structured Approach to Decision Making. Report of the Workshop organised by the CMR International Institute for Regulatory Science at the Georgetown Inn, Washington D.C., USA 13-14 June 2005. Available from URL: http://www.mhra.gov.uk/home/groups/espolicy/documents/websiteresources/con028325.pdf. 2005.

76 . Walker S., Cone M. Benefit-Risk Assessment: The Development of a Model for Benefit-Risk Assessment of Medicines Based on Multi-Criteria Decision Analysis London, UK: 29-30 March 2004. CMR International Institute for Regulatory Science. Available from URL: http://www.cmr.org/. 2004.

77 . Walker S., Cone M. Measuring Benefit and Balancing Risk: Strategies for the benefit-risk assessment of new medicines in a risk-averse environment Washington, DC, US: 19-20 June 2008. Report of the Workshop organised by the CMR International Institute for Regulatory Science. Available from URL: http://www.cmr.org. 2008.

78 . Liberti L, Cone M. Strategies for Communicating Benefit-Risk to Decision Makers: Explaining Methods, Findings and Conclusions Through a Common Approach Washington, DC, US: 17-19 June 2009. Report of the Workshop organised by the CMR International Institute for Regulatory Science. Available from URL: http://www.cmr.org/workshops. 2009.

79 . Walker S, Liberti L, McAuslane N. Refining the benefit-risk framework for the assessment of medicines: valuing and weighting benefit and risk parameters. *Clin Pharmacol Ther*. 2011 Feb;89(2):179-82.

80 . Value Tree Analysis. Available online from URL: http://www.mcda.hut.fi/value_tree/theory/theory.pdf. 2002 Apr 30.

81 . Liberti L, McAuslane N, Walker S. Progress on the development of a benefit/risk framework for evaluating medicines. *Regulatory Focus*. 2010 Mar;1-6.

82 . Coplan PM, Noel RA, Levitan BS, Ferguson J, Mussen F. Development of a framework for enhancing the transparency, reproducibility and communication of the benefit-risk balance of medicines. *Clin Pharmacol Ther*. 2011 Feb;89(2):312-5.

83 . Ashby D. Bayesian statistics in medicine: a 25 year review. *Stat Med*. 2006 Nov 15;25(21):3589-631.

84 . Ashby D, Smith AF. Evidence-based medicine as Bayesian decision-making. *Stat Med*. 2000 Dec 15;19(23):3291-305.

85 . Naimark D, Krahn MD, Naglie G, Redelmeier DA, Detsky AS. Primer on medical decision analysis: Part 5--Working with Markov processes. *Med Decis Making*. 1997 Apr;17(2):152-9.

86 . Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. *Med Decis Making*. 1993 Oct;13(4):322-38.

87 . ICH Guidelines. Available from URL: http://www.ich.org/cache/compo/276-254-1.html. 2010.
Ref Type: Online Source

88 . Significant withdrawals. Available from URL: http://en.wikipedia.org/wiki/List_of_withdrawn_drugs. 2011.
Ref Type: Online Source

89 . Benefit-risk methodology project - work package 3 report: Field tests. EMA/718294/2011 [online]. Available from URL:http://www.ema.europa.eu/docs/en_GB/document_library/Report/2011/09/WC500112088.pdf. European Medicines Agency (EMA); 2011.

90 . Walker S, McAuslane N, Liberti L, Salek S. Measuring benefit and balancing risk: strategies for the benefit-risk assessment of new medicines in a risk-averse environment. *Clin Pharmacol Ther*. 2009 Mar;85(3):241-6.

91 . PROTECT Home page. URL: http://www.imi-protect.eu/objectives.html. 2011.
Ref Type: Online Source

92 . Dr. Christine E. Hallgreen, Post. Doc. PROTECT Work Package 5. 2011.
Ref Type: Personal Communication

93 . Lindgren BR, Wielinski CL, Finkelstein SM, Warwick WJ. Contrasting clinical and statistical significance within the research setting. *Pediatr Pulmonol*. 1993 Dec;16(6):336-40.

94 . Watkins PB. Drug safety sciences and the bottleneck in drug development. *Clin Pharmacol Ther*. 2011 Jun;89(6):788-90.

95 . Mullard A. 2010 FDA drug approvals. *Nat Rev Drug Discov*. 2011 Feb;10(2):82-5.

96 . European Medicines Agency: Clinical efficacy and safety guidelines introduction. Available from URL: http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_con

tent_000085.jsp&murl=menus/regulations/regulations.jsp&mid=WC0b01ac058002754949&jsenabled=true. 2011.
Ref Type: Online Source

97 . The Licensing Department at the Danish Medicines Agency. 2011.
Ref Type: Personal Communication

98 . Fienberg SE. A Brief History of Statistics in Three and One-Half Chapters: A Review Essay. *Statistical Science*. 1992;7(2):208-25.

99 . McCloskey DN, Ziliak ST. The Unreasonable Ineffectiveness of Fisherian "Tests" in Biology and Especially Medicine. *Biological Theory*. 2009;4(1):44-53.

100 . Greenstein G. Clinical versus statistical significance as they relate to the efficacy of periodontal therapy. *J Am Dent Assoc*. 2003 May;134(5):583-91.

101 . Hollon SD, Flick SN. On the Meaning and Methods of Clinical Significance. *Behavioral Assessment*. 1988;10:197-206.

102 . Lefort SM. The statistical versus clinical significance debate. *Image J Nurs Sch*. 1993;25(1):57-62.

103 . Kant I. Critique of pure reason. Cambridge University Press; 1998.

104 . Søren Kierkegaards Skrifter, bind 7 : Afsluttende uvidenskabelig Efterskrift. Gads Forlag, Denmark; 2002.

105 . Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol*. 1991 Feb;59(1):12-9.

106 . Jacobson NS, Folette WC, Revenstorf D. Psychotherapy Outcome Research: Methods for Reporting Variability and Evaluating Clinical Significance. *Behavior Therapy*. 1984;15:336-52.

107 . Atkins DC, Bedics JD, McGlinchey JB, Beauchaine TP. Assessing clinical significance: does it matter which method we use? *J Consult Clin Psychol*. 2005 Oct;73(5):982-9.

108 . Campbell A. Clinical Significance in Real World Settings. *ANZJFT*. 2008;29(2):107-10.

109 . Kingman A. Statistical vs clinical significance in product testing: can they be designed to satisfy equivalence? *J Public Health Dent*. 1992;52(6):353-60.

110 . Hujoel PP, Armitage GC, Garcia RI. A perspective on clinical significance. *J Periodontol*. 2000 Sep;71(9):1515-8.

111 . Killoy WJ. The clinical significance of local chemotherapies. *J Clin Periodontol*. 2002 May;29 Suppl 2:22-9.

112 . Grinstead CM, Snell JL. Sums of Random Variables. Introduction to Probability. 2 ed. American Mathematical Society; 1997. p. 285.

113 . Sarac SB, Rasmussen CH, Rasmussen MA, Hallgreen CE, Søeborg T, Colding-Jørgensen M, et al. A Comprehensive Approach to Benefit-Risk Assessment in Drug Development. Basic & Clinical Pharmacology & Toxicology. In press 2012.

114 . Giacchetti S, Perpoint B, Zidani R, Le BN, Faggiuolo R, Focan C, et al. Phase III multicenter randomized trial of oxaliplatin added to chronomodulated fluorouracil-leucovorin as first-line treatment of metastatic colorectal cancer. *J Clin Oncol*. 2000 Jan;18(1):136-47.

115 . Clopper C, Pearson E. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*. 1934;26(4):404-13.

116 . Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.

117 . Afzal S, Gusella M, Jensen SA, Vainer B, Vogel U, Andersen JT, et al. The association of polymorphisms in 5-fluorouracil metabolism genes with outcome in adjuvant treatment of colorectal cancer. *Pharmacogenomics*. 2011 Sep;12(9):1257-67.

118 . Sarac SB, Rasmussen CH, Afzal S, Thirstrup S, Colding-Jørgensen M, Poulsen HE, et al. Data-driven assessment of the association of polymorphisms in 5-Fluorouracil metabolism genes with outcome in adjuvant treatment of colorectal cancer. Basic & Clinical Pharmacology & Toxicology. In press 2012.

119 . Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin*. 2011 Mar;61(2):69-90.

120 . Afzal S, Jensen SA, Vainer B, Vogel U, Matsen JP, Sorensen JB, et al. MTHFR polymorphisms and 5-FU-based adjuvant chemotherapy in colorectal cancer. *Ann Oncol*. 2009 Oct;20(10):1660-6.

121 . Afzal S, Gusella M, Vainer B, Vogel UB, Andersen JT, Broedbaek K, et al. Combinations of polymorphisms in genes involved in the 5-Fluorouracil metabolism pathway are associated with gastrointestinal toxicity in chemotherapy-treated colorectal cancer patients. *Clin Cancer Res*. 2011 Jun 1;17(11):3822-9.

122 . EMA. EPAR on Ketek. Available online: http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/000354/human_med_000873.jsp&murl=menus/medicines/medicines.jsp&jsenabled=true. 2009.
Ref Type: Online Source

Index